COMPLEX SURVEY AND DESIGN EFFECT WITH RAO-SCOTT CORRECTION AND LOG-LINEAR ANALYSIS

A Thesis Submitted to the College of
Graduate and Postdoctoral Studies
In Partial Fulfillment of Master of Science
In the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon

By

SANJEEV RIJAL

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis/dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

DISCLAIMER

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
142-106 Wiggins Rd.
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5E6
Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
105 Administration Place
Saskatoon, Saskatchewan S7N 5A2
Canada

ABSTRACT

Collection of data through sample surveys involves a wide range of techniques and procedures. Data is collected with the priority of maximum accuracy of information and with minimum cost and effort. Various sampling techniques are used to achieve the objective of accuracy and low cost.

When the data is collected from a complex surveys techniques for the analysis of categorical data (e.g. the chi-squared test for association between pairs of variable) have to be modified from the procedures used when the sample design is assumed to be a simple random sample. When the data is acquired using various sampling techniques in a complex design, there is a sample design effect on whatever analysis is used. Rao-Scott showed the effect of design on the tests of fitness, homogeneity and independence. The Log-linear model is used for analysis of higher dimension categorical data. The modifications of this analysis for complex survey designs were also proposed by Rao-Scott.

Simultaneous Test Procedure, another method, to test the homogeneity between multiple categories can also be linked with log-likelihood ratio statistics.

The fundamental concepts of data collection are explained with examples. The basic concepts of test procedure along with ways to get log-linear models are discussed leading to the multi-dimension with general log-linear model. The examples following the concepts show the validity in calculation.

ACKNOWLEDGEMENTS

I am sincerely grateful to my supervisor Dr. William H. Laverty for his continuous support, guidance and directions on my research. I extend my gratitude to the department of mathematics and statistics of University of Saskatchewan, Dr. Ebrahim Samei and all other who created encouraging and providing academic environment to carry on with my research.

TABLE OF CONTENTS

Page

1. INTRODUCTION

It has been an ever-going pursuit and attempt of humanity to understand the scientific and mathematical design and patterns underlying natural phenomena. The way things are set and settled around us and in nature has some mathematical significance to it. From the very early days of human civilization, we have tried to understand those secrets of presence, coexistence, and settings between and of things so that we can figure out ways to gain and maximize benefits. Numbers as counts and data in nature are laid out in some natural way, which tells and carries certain information on them. Mathematicians from the very early days have been working on it and have and still are coming up with various findings and results from those numbers or data. The first and foremost thing to be done before we work on the data is to collect the data itself. And the most common way to collect or obtain data on some population is from sample surveys, which can vary from a simple one to a multistage complex one. When collecting data from a large population on various factors, the objective is to collect the most accurate information at the least cost. The survey involving large population frequently ends up being a complex one. Complex surveys are widely used in various fields including agricultural, pharmaceutical, marketing, product development etc. Complex surveys are collections of data on interested factors or variables using some kind of complex sampling design. Complex surveys usually deals with large sample size with highly significant research goal or objective. Samples from complex surveys are probability samples that don't fit the conventions of Simple Random Samples, that is, the sample member doesn't have an equal probability of being selected. Complex surveys involve many features like stratification, clustering, oversampling, non-replacement sampling, finite population and multistage sampling, which results in its complexities. The reasons for doing complex surveys vary with its objective. The first most important one is, it is less costly

and the sampling method gives improved sample estimate, secondly, it allows us to analyze several levels of the sample data (as in multistage sampling).

In this thesis, I am talking about the analysis of complex surveys containing cross-classified categorical data. Among many known general tests and techniques to analyze such data, the chi-squared test for association of two variables is the most common one. For analysis of data containing three or more variables, a more appropriate way of analyzing is the log-linear analysis. I have discussed various ways and techniques of collecting data and mathematical ways of calculating statistical information and their interpretation in a simple possible way. I have used the concept and examples published in the papers by Rao and Scott (1979), Rao and Thomas (1987, 1988), K.R. Gabriel (1966, 1969) and many others listed in reference discussing the process of analyzing complex survey data. When working with big complex survey data the origin or more precisely the process involved in obtaining those data plays a significant role in the analytic result. Obtained data in complex surveys goes through different ways of sampling techniques and various stages which turn out to yield data which are quite different from the data used for standard analysis with simple random sampling assumption. Particularly collecting data through stratification and clustering with multiple stages has some effects on the nature of data collected and hence should be handled differently than the standard simple random sampled data. This effect changes the nature of the distribution. Certain statistics based on large samples tend to have chi-squared distribution under SRS (Simple Random Sampling) techniques but when the clustering techniques are involved it distorts the distribution. Using standard analysis, assuming SRS can sometimes result in highly misleading conclusions for the complex surveys. Rao and Scott's papers formulated the first and second order correction to the statistics of complex survey data to cope with the effect of sampling. I have tried to explain the theory behind those corrections with examples from Rao and Thomas (1988) paper and verified it with other sampled data as well. This paper also discusses the concept of log-linear model for three or higher dimension table of complex survey data. The log-linear model concept discussed herein is to explain how it is derived for a two-way table and then for three ways and for higher dimensions as well. The purpose of the detailed explanation is to understand the general in Rao and Thomas (1988) for higher dimension, which is generally the case in complex survey data.

I have also discussed the validity of small sample test under big complex surveys with simultaneous test procedure (STP). K.R. Gabriel has shown STP to test the homogeneity of the

different sets of categorical data within a complex survey. I have used the concept with the example in this thesis.

2. SAMPLING TECHNIQUES

Sampling of the population is carried out to study, analysis, estimate or collect information on that population for various purposes. Basically, sampling is done using two different techniques: probability sampling techniques and non-probability sampling techniques. Probability sampling involves some kind of random selection of the sample with some (equal or un-equal) probability of being selected whereas non-probability sampling does not involve random selection.

There are various probability sampling techniques used by the statisticians and researchers in the field. The technique selected for the purpose depends on the characteristic of the population sample. Following are the major sampling techniques,

- Simple Random Sampling
- Stratified Sampling
- Cluster Sampling
- Multistage Sampling
- Unequal-Probability Sampling

2.1 *Simple Random Sampling (SRS)*

It is the basic sampling technique. This sampling technique is used if the chance or probability of selecting, picking or including sample or unit from the population for the study or analysis is equal. SRS requires minimal prior knowledge of the population and the frame (list of

the sampling units). It is appropriate if the collection of data can be done effectively without much cost on randomly distributed items.

Simple random sample is produced in various steps. First, the population to be studied is defined. Like the population of a country, county, town, University, School etc. is considered. Then the sample size is determined. This depends on the budget, manpower and other required facilities available for the study. Then the data on the population to be studied is obtained. Each unit in the sample is then assigned random numbers and desired size of the sample is selected out of the entire population. And then the sample hence selected is analyzed using statistical methods.

The benefit of using SRS reduces the potential of human biases in the process of selecting the sample. The sample hence selected will represent the whole population. As a result, the statistical information drawn from the sample can be used to make a valid general conclusion for the entire population. SRS is cost effective and simple making it the preferred way of doing study or analysis of the population where possible.

One major requirement for SRS is the availability of the list of the population. Obtaining the list sometime requires a lot of financial, administrative or legal efforts. Beside that collecting population information where possible may require a lot of effort, time and determination based on a geological, economical and social position of the population.

For example let's consider a population with a total number of elements $N = 5$ and the population itself $\{3, 6, 9, 12, 15\}$. From the population the mean $\mu = 9$ and the population variance $\sigma^2 = 18$. Simple random sample of size $n = 3$ is picked from the above population without replacing. Then,

Possible Samples	Probability of picking the sample	Mean of the sample	Variance of the sample
{3, 6, 9}	1/10	6	9
{3, 6, 12}	1/10	7	21
{3, 6, 15}	1/10	8	39
{6, 9, 12}	1/10	9	9
{6, 9, 15}	1/10	10	21
{9, 12, 15}	1/10	12	9
{3, 12, 15}	1/10	10	39
{3, 9, 12}	1/10	8	21
{3, 9, 15}	1/10	9	36
{6, 12, 15}	1/10	11	21

From the above table, we can calculate expected value of sample means. The expected value of sample means is given by,

$$E(\bar{x}) = \sum_{k=1}^{K} \overline{x_k}\, p(\overline{x_k}) \text{ ……..(2.1.1)}$$

Here,

$$E(\bar{x}) = \sum_{k=1}^{10} \overline{x_k}\, p(\overline{x_k})$$

$$= \frac{1}{10}(6 + 7 + 8 + 9 + 10 + 12 + 10 + 8 + 9 + 11)$$

$$= \frac{1}{10}(90) = 9$$

That is, $E(\bar{x}) = \mu = 9$. And the variance of the sample means is given by,

$$V(\bar{x}) = E(\bar{x}^2) - (E(\bar{x}))^2 \text{.....(2.1.2)}$$

$$= E(\bar{x}^2) - (9)^2$$

where $E(\bar{x}^2) = \sum_{k=1}^{K} \overline{x_k}^2 \, p\left(\overline{x_k}^2\right) = \sum_{k=1}^{10} \overline{x_k}^2 \, p\left(\overline{x_k}^2\right)$ .....(2.1.3)

$$= \frac{1}{10}(6^2 + 7^2 + 8^2 + 9^2 + 10^2 + 12^2 + 10^2 + 8^2 + 8^2 + 11^2)$$

$$= \frac{1}{10}(840) = 84$$

Then,

$$V(\bar{x}) = E(\bar{x}^2) - (9)^2 = 84 - 81 = 3$$

and ,

$$V(\bar{x}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \text{.....(2.1.4)}$$

$$= \frac{18}{3}\left(\frac{5-3}{5-1}\right) = 3$$

We get expected values of the sample variance by,

$$E(s^2) = \sum_{k=1}^{K} {s_k}^2 \, p({s_k}^2) = \sum_{k=1}^{10} {s_k}^2 \, p({s_k}^2) \text{.....(2.1.5)}$$

$$= \frac{1}{10}(9 + 21 + 39 + 9 + 21 + 9 + 39 + 21 + 36 + 21)$$

$$= \frac{1}{10}(225) = 22.5$$

On the other hand, from population variance, we get the same expected value as,

$$E(s^2) = \left(\frac{N}{N-1}\right)\sigma^2 \text{.....(2.1.6)}$$

$$= \left(\frac{5}{5-1}\right)(18)$$

$$= \frac{5}{4}(18) = 22.5$$

Simple random sampling is the base for all other sampling techniques if randomness and equal probability sampling is to be considered. All other techniques, although uses various methods to allocate the frame, but at last uses SRS to draw the observation or sample. For SRS the whole population is considered and it is sometime not economically and geographically possible. It sometimes requires more manpower to collect SRS. Since the whole population is considered no classification error is encountered. SRS is known for its simplicity, which can be conducted without the prior knowledge of population sample.

2.2. *Stratified Sampling*

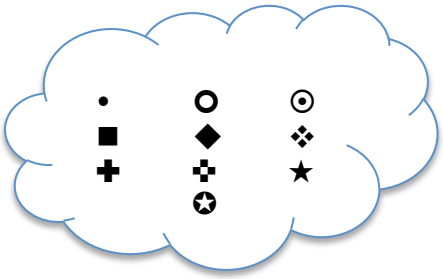
"A stratified sampling is one obtained by separating the population elements into non-overlapping groups called strata and then selecting a simple random sample from each stratum" (Scheaffer, R. L., W. Mendenhall and R. Lyman Ott, 1979, p.59)

Stratified sampling is used if there is a specific subgroup in the population, which is to be given more importance. Stratification may result in a smaller error of estimation particularly, if measurements within strata are homogeneous. This method of sampling is cost effective as the sampling can be divided into sub-groups and requires handling less coverage area. It also provides separate estimates of the elements, sub-groups or stratum under the populations without additional sampling. The size of the sample to be dealt with in this sampling is small, reducing the cost. But the administrative effort in this sampling method is high as there are many subgroups to be considered and the data should be collected separately for each subgroup.

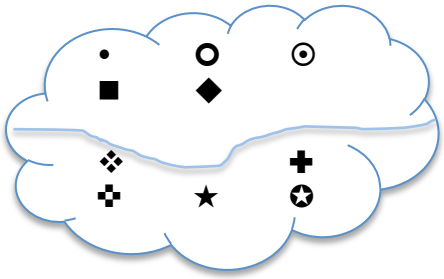
To select a stratified sample, an entire population is divided into various strata or groups according to the elements to be studied. And from those strata, a SRS is drawn for study or analysis. But we have to keep in mind that stratified sample is not a simple random sample.

2.2.1*Illustrating difference between SRS and Stratified sampling*:

Let's consider a group of ten different symbols.



Group A



Group B

Let's pick a simple random sample of four symbols from Group A. The probability of the symbol ✪ being in that sample is P(✪)= $\frac{4}{10}$. And the probability of picking a sample group {✪, ❖, ✚, ✣} out of total possible $\binom{10}{4} = 210$ samples is P(✪, ❖, ✚, ✣)= $\frac{1}{210}$.

In Group B, the symbols are divided into two strata. We pick two from the upper stratum and two from the lower stratum to get a group of four symbols. The probability of symbol ✪ being in the sample is still P(✪)= $\frac{4}{10}$. But the probability of picking a sample group {✪, ❖, ✚, ✣} is P(✪, ❖, ✚, ✣)=0. This shows that the probability of picking a sample is not same in stratified random sample and hence in not the same as simple random sample.

As an example for Stratified sampling, Let's consider data from four different strata of agricultural grain production field with sample production and the following information,

Strata	Total area in acres (N_k)	Sample size (n_k)	Sample production data bushels/acre (x_k)	Mean ($\bar{x}_k$)	Sample Variance (σ_k)
1	3000	5	{54,58,52,62,60}	57.2	17.2
2	4000	6	{42,52,46,48,50,44}	47	14
3	2500	4	{56, 62, 54, 60}	58	13.33
4	7000	8	{44, 46, 45, 44, 56, 52, 47, 55}	48.62	24.55

In general, if we have K strata of size N_i with $\sum_{k=1}^{K} N_k = N$ and sample size on n_i from each stratum with $\sum_{k=1}^{K} n_k = n$ then, the estimator of the total is given by

$$\hat{T} = \sum_{k=1}^{K} N_k \bar{x}_k \text{.....(2.2.1)}$$

Here

$$\hat{T} = 3000(57.2) + 4000(47) + 2500(58) + 7000(48.62) = 844940$$

where $\bar{x}_k$ is the mean of sample n_k. And the estimator of the mean is given by

$$\hat{\mu} = \bar{x} = \sum_{k=1}^{K} \frac{N_k}{N} \bar{x}_k \text{.....(2.2.2)}$$

$$\hat{\mu} = \bar{x} = \sum_{k=1}^{K} W_k \bar{x}_k$$

where W_k is the weights and given by population proportions, i.e. $W_k = \frac{N_k}{N}$.

Here, for the above example,

$$\hat{\mu} = \bar{x} = \sum_{k=1}^{K} \frac{N_k}{N} \bar{x}_k = \frac{1}{N}\left(\sum_{k=1}^{K} N_k \bar{x}_k\right)$$

$$= \frac{1}{16500}(844940) = 51.21$$

The variance is then given by

$$V(\bar{x}) = \sum_{k=1}^{K} W_k^2 V(\bar{x}_k) \ldots\ldots\ldots\ldots\ldots\ldots (2.2.3)$$

$$V(\bar{x}) = \sum_{k=1}^{K} W_k^2 \frac{\sigma_k^2}{n_k} = \sum_{k=1}^{K} \frac{W_k^2 \sigma_k^2}{n w_k}$$

where sample weight $w_k = \frac{n_k}{n}$. For the above example,

$$V(\bar{x}) = \sum_{k=1}^{K} W_k^2 \frac{\sigma_k^2}{n_k} = \left(\frac{30}{165}\right)^2 \frac{17.2^2}{5} + \left(\frac{40}{165}\right)^2 \frac{14^2}{6} + \left(\frac{25}{165}\right)^2 \frac{13.33^2}{4} + \left(\frac{70}{165}\right)^2 \frac{24.55^2}{8}$$

$$= 1.95 + 1.92 + 1.02 + 13.56 = 18.45$$

Let the approximation of the variance of the production from the four strata based on previous records be $\sigma_1 = 10$ bushels/acre, $\sigma_2 = 12$ bushels/acre, $\sigma_3 = 8$ bushels/acre, $\sigma_4 = 16$ bushels/acre. To estimate the mean yield in bushels per acre for the four strata considering a margin of error 5 bushels/acre, we find n and $n_k's$.

Here we have

$$\frac{n_k}{n} = \frac{N_k}{N} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.2.4)$$

$$n_1 = \frac{30}{165} n \, , n_2 = \frac{40}{165} n \, , n_3 = \frac{25}{165} n \; and \; n_4 = \frac{70}{165} n$$

$$W_1 = w_1 = \frac{30}{165}, W_2 = w_2 = \frac{40}{165}, W_3 = w_3 = \frac{25}{165} \; and \; W_4 = w_4 = \frac{70}{165}$$

Now for proportion allocation,

$$n = \frac{4}{M^2}\left[\sum_{k=1}^{K}\frac{W_k^2\sigma_k^2}{W_k}\right]\text{.....(2.2.5)}$$

where M is the desired margin of error.

$$= \frac{4}{5^2}\left[\frac{\left(30/165\right)^2 10^2}{\left(30/165\right)} + \frac{\left(40/165\right)^2 12^2}{\left(40/165\right)} + \frac{\left(25/165\right)^2 8^2}{\left(25/165\right)} + \frac{\left(70/165\right)^2 16^2}{\left(70/165\right)}\right]$$

$$= \frac{4}{25}(18.182 + 34.91 + 9.697 + 108.606)$$

$$= \frac{4}{25}(171.395) = \frac{685.58}{25}$$

$$= 27.42 \approx 27$$

Then,

$$n_1 = \frac{30}{165}(27) = 4.91 \approx 5$$

$$n_2 = \frac{40}{165}(27) = 6.54 \approx 7$$

$$n_3 = \frac{25}{165}(27) = 4.09 \approx 4$$

$$n_4 = \frac{70}{165}(27) = 11.45 \approx 11$$

For optimal allocation of sample size to minimize the variance (Neyman allocation, developed by the statistician Jerzy Neyman)

$$n_k = \left(\frac{W_k\sigma_k}{\sum_{k=1}^{K} W_k\sigma_k}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.2.6)$$

$$\sum_{k=1}^{4} W_k\sigma_k = W_1\sigma_1 + W_2\sigma_2 + W_3\sigma_3 + W_4\sigma_4 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.2.7)$$

$$= \frac{30}{165}(10) + \frac{40}{165}(12) + \frac{25}{165}(8) + \frac{70}{165}(16)$$

$$= 1.82 + 2.91 + 1.21 + 6.79 = 12.73$$

$$n_1 = n\left(\frac{W_1\sigma_1}{12.73}\right) = n\left(\frac{1.82}{12.73}\right) = 0.143n$$

$$n_2 = n\left(\frac{W_2\sigma_2}{12.73}\right) = n\left(\frac{2.91}{12.73}\right) = 0.228n$$

$$n_3 = n\left(\frac{W_3\sigma_3}{12.73}\right) = n\left(\frac{1.21}{12.73}\right) = 0.095n$$

$$n_4 = n\left(\frac{W_4\sigma_4}{12.73}\right) = n\left(\frac{6.79}{12.73}\right) = 0.533n$$

$$n = \frac{4}{M^2}\left[\sum_{k=1}^{K} \frac{W_k^2\sigma_k^2}{w_k}\right]$$

$$n = \frac{4}{5^2}\left[\frac{W_1^2\sigma_1^2}{w_1} + \frac{W_2^2\sigma_2^2}{w_2} + \frac{W_3^2\sigma_3^2}{w_3} + \frac{W_4^2\sigma_4^2}{w_4}\right]$$

$$= \frac{4}{25}\left[\frac{\left({}^{30}/_{165}\right)^2 10^2}{0.143} + \frac{\left({}^{40}/_{165}\right)^2 12^2}{0.228} + \frac{\left({}^{25}/_{165}\right)^2 8^2}{0.095} + \frac{\left({}^{70}/_{165}\right)^2 16^2}{0.533}\right]$$

$$= \frac{4}{25}(23.12 + 37.12 + 15.46 + 86.44)$$

$$= \frac{4}{25}(162.14) = \frac{648.56}{25} = 25.94 \approx 26$$

So, if $n = 26$ then,

$$n_1 = (0.143)(26) = 3.718 \approx 4$$

$$n_2 = (0.228)(26) = 5.92 \approx 6$$

$$n_3 = (0.095)(26) = 2.47 \approx 2$$

$$n_4 = (0.533)(26) = 13.858 \approx 14$$

Comparing optimal allocation of sample sizes to the proportional allocation we can see the sample size decreased for strata 1, 2 and 3 and increase in size for stratum 4. This is so as stratum 4 has a bigger size with greater variance and allocating larger sample size minimizes the variance.

Stratified sampling is best suited if the area or region to be sampled has various groups in it. The variance within the strata will be less and among the strata will be more. It yields smaller estimation errors. If the sampling frame does not have homogeneous group pocket that can be separated as strata then the stratified sampling in not appropriate.

2.3.*Cluster Sampling*

This sampling technique is used if the population to be studied is heterogeneous. Cluster sampling is a sampling technique in which the total population is divided into clusters or groups and simple random sample is drawn from these clusters. This is then also called 'two-stage' design as it involves cluster and simple random sampling.

This sampling technique reduces the administration, travel and listing cost given the frame is provided. It also gives an accurate result for most of the variation within the group. It is effective when dealing with big population. As a disadvantage of this sampling method, it produces higher sampling error and biased samples.

Dividing the entire population into different clusters or subgroups, cluster sampling procedure is carried out. These clusters are heterogeneous, containing all the elements of the population. And either all cluster is considered or SRS of clusters is used and SRS of elements is selected from each cluster.

Let's consider a city area divided into 345 different clusters. Out of the total 345 clusters, only 25 different clusters are randomly selected as simple random sample. And the survey is done at every household in those sampled clusters. The data on the amount of money spend on health care annually by each household are presented in the table below as a sum of respective clusters.

Clusters ($k's$)	No. of households is the cluster (h_k)	Total per cluster health expenditure (x_k)
1	5	11000
2	6	8000
3	2	17000
4	3	10000
5	8	9000
6	5	13000
7	7	15000
8	6	8000
9	6	10000
10	5	13000
11	4	8000
12	16	24000
13	8	19000
14	8	8000
15	3	9000
16	7	8000
17	8	6000
18	6	10000
19	4	8000
20	5	9000
21	5	4000
22	6	6000
23	3	10000
24	9	11000
25	10	10000
	155	264000

In the above table, the total number of clusters, $N = 345$. Total number of sampled clusters, $n = 25$. Total number of households in sampled clusters, $\sum_{k=1}^{n} h_k = 155$. Total expenditure of all households in sampled clusters, $\sum_{k=1}^{n} x_k = 264\,000$.

Now, the best estimate of the mean expenditure of the household μ is,

$$r = \frac{\sum_{k=1}^{n} x_k}{\sum_{k=1}^{n} h_k} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3.1)$$

$$= \frac{264000}{155} = 1703$$

The per-household expenditure in health annually is \$1703. Here r is known as population ratio. Here the total number of households, K, in all 345 clusters is not known so we estimate $\overline{H}$ using

$$\overline{H} = \frac{\sum_{k=1}^{n} h_k}{n} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3.2)$$

$$= \frac{155}{25} = 6.2$$

Now the estimate of variance of population mean $\bar{x}$ also known as population ratio r is given by,

$$\hat{V}(\bar{x}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{\overline{H}^2}\right)\left(\frac{s_r^2}{n}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3.3)$$

where s_r^2 is the estimated variance of r and is given by,

$$s_r^2 = \frac{\sum_{k=1}^{n} (x_k - r k_k)^2}{n-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3.4)$$

For the above problem $s_r^2 = 24539176.29$

$$\hat{V}(\bar{x}) = \left(\frac{345 - 25}{345}\right)\left(\frac{1}{6.2^2}\right)\left(\frac{24539176.29}{25}\right)$$

$$= 23684.68$$

And the error estimate is

$$2\sqrt{\hat{V}(\bar{x})} = 2(153.9) = 307.8$$

Hence, the estimate of μ with the bound is

$$\hat{\mu} = \bar{x} \pm 2\sqrt{\hat{V}(\bar{x})} \quad \text{.....(2.3.6)}$$

$$= 1703 \pm 308$$

2.4. ***Multistage Sampling***

Multistage sampling is a 'higher' level of cluster sampling. As the name says, this sampling technique is carried out in multiple staged using relatively smaller and smaller sampling unit at every stage. Multistage sampling involves inflation or deflation of frame within the subgroup. For example, Let's consider the survey done by Postal Service to get information on the flow of snail mails in the city. For the purpose, the city is divided into collection areas and selecting some of the collection areas (1st stage). The selected collection areas are then divided into blocks and blocks are selected from each collection areas (2nd stage). Then houses in the selected blocks are selected (3rd stage). The big city population is now reduced to small sample including houses from every collection areas in the city.

This type of sampling methods can vary from a simple to complex multistage sampling. It seems like stratified sampling and cluster sampling carries similarities with multistage sampling but is subsequently different. In cluster sampling selected clusters are studied and in stratified sampling, SRS is selected from all the strata. The benefit of conducting multistage

sampling is its cost effective and efficient and more accurate then cluster sampling for same sample size.

2.5. *Unequal Probability Sampling*

The sampling techniques, discussed above, are used if the possibilities of selection of units in the sample have equal probability. Sometimes this isn't the case. In some cases, the possibilities of selecting units in the sample have unequal probabilities. Unequal probability sampling technique is used when some units in the population have higher probabilities of being selected from others. This sampling technique is used in two different ways:

i) Selection Probability and

ii) Inclusion Probability

The Hansen-Hurwitz estimator for sampling with replacement (the selection probabilities do not change after every draw) is based on selection probability. And Horvitz-Thompson estimator for sampling with or without replacement is based on inclusion probability.

As an example let's consider the production of three different factories and their probabilities of being selected as a supplier to the consumer market.

Factory	1	2	3
Selection prob.(p_k)	0.2	0.5	0.3
Production (x_k)	30	90	55

From above Let's pick a sample of two possible sets of factories with replacement, that is $n = 2$ and $N = 3$ (factories).

Samples(s)	$p(s)$	x's
(1,1)	(0.2)(0.2) = 0.04	(30,30)
(1,2)	(0.2)(0.5) = 0.1	(30,90)
(2,1)	(0.5)(0.2) = 0.1	(90,30)
(1,3)	(0.2)(0.3) = 0.06	(30,55)
(3,1)	(0.3)(0.2) = 0.06	(55,30)
(2,2)	(0.5)(0.5) = 0.25	(90,90)
(2,3)	(0.5)(0.3) = 0.15	(90,55)
(3,2)	(0.3)(0.5) = 0.15	(55,90)
(3,3)	(0.3)(0.3) = 0.09	(55,55)

The Hansen-Hurwitz estimator in general is given by,

$$H_U(s) = \frac{1}{n}\sum_{k=1}^{n}\frac{x_k}{p_k} \quad \text{.....(2.5.1)}$$

In above example Hansen-Hurwitz estimator for the sample (1,1) is

$$H_U(1,1) = \frac{1}{2}\left(\frac{x_1}{p_1} + \frac{x_1}{p_1}\right)$$

$$= \frac{1}{2}\left(\frac{30}{0.2} + \frac{30}{0.2}\right) = 150$$

and similarly for samples (1,2), (2,1), (1.3), (3,1), (2,2), (2,3), (3,2) and (3,3) are

$$H_U(1,2) = \frac{1}{2}\left(\frac{x_1}{p_1} + \frac{x_2}{p_2}\right) = \frac{1}{2}\left(\frac{30}{0.2} + \frac{90}{0.5}\right) = 115$$

$$H_U(2,1) = \frac{1}{2}\left(\frac{x_2}{p_2} + \frac{x_1}{p_1}\right) = \frac{1}{2}\left(\frac{90}{0.5} + \frac{30}{0.2}\right) = 115$$

$$H_U(1,3) = \frac{1}{2}\left(\frac{x_1}{p_1} + \frac{x_3}{p_3}\right) = \frac{1}{2}\left(\frac{30}{0.2} + \frac{55}{0.3}\right) = 166.67$$

$$H_U(3,1) = \frac{1}{2}\left(\frac{x_3}{p_3} + \frac{x_1}{p_1}\right) = \frac{1}{2}\left(\frac{55}{0.3} + \frac{30}{0.5}\right) = 166.67$$

$$H_U(2,2) = \frac{1}{2}\left(\frac{x_2}{p_2} + \frac{x_2}{p_2}\right) = \frac{1}{2}\left(\frac{90}{0.5} + \frac{90}{0.5}\right) = 180$$

$$H_U(2,3) = \frac{1}{2}\left(\frac{x_2}{p_2} + \frac{x_3}{p_3}\right) = \frac{1}{2}\left(\frac{90}{0.5} + \frac{55}{0.3}\right) = 181.67$$

$$H_U(3,2) = \frac{1}{2}\left(\frac{x_3}{p_3} + \frac{x_2}{p_2}\right) = \frac{1}{2}\left(\frac{55}{0.3} + \frac{90}{0.5}\right) = 181.67$$

$$H_U(3,3) = \frac{1}{2}\left(\frac{x_3}{p_3} + \frac{x_3}{p_3}\right) = \frac{1}{2}\left(\frac{55}{0.3} + \frac{55}{0.3}\right) = 183.33$$

Mean of $H_U = E(H_U) = \sum p(s)H_U(s)$…….(2.5.2)

$$= (0.04)(150) + (0.1)(115) + (0.1)(115) + (0.06)(166.67)$$
$$+(0.06)(166.67) + (0.25)(180) + (0.15)(181.67) + (0.15)(181.67) + (0.09)(183.33)$$
$$= 165$$

and the variance, $Var(H_U) = E[H_U - mean\ of\ H_U]^2$…….(2.5.3)

where $E(H_U) = \sum p(s)q(s)$, $q(s) = (H_U - mean\ of\ H_U)^2$ ……..……..(2.5.4)

Then,

$$Var(H_U) = (0.04)(150 - 165)^2 + (0.1)(115 - 165)^2 + (0.1)(115 - 165)^2$$
$$+ (0.06)(166.67 - 165)^2 + (0.06)(166.67 - 165)^2$$

$$+(0.25)(180-165)^2+(0.15)(181.67-165)^2+(0.15)(181.67-165)^2$$
$$+(0.09)(183.33-165)^2=679.19$$

Next, for the Horvitz-Thompson estimator, the probabilities that factory 1, 2 and 3 will be in the sample is as follows respectively,

$$H_{T1}=\ p_1(1,1)+\ p_2(1,2)+\ p_3(2,1)+\ p_4(1,3)\ +\ p_5(3,1)$$
$$H_{T1}=\ 0.04+0.1+0.1+0.06+0.06=0.36$$

$$H_{T2}=\ p_2(1,2)+\ p_3(2,1)+\ p_6(2,2)+\ p_7(2,3)\ +\ p_8(3,2)$$
$$H_{T2}=\ 0.1+0.1+0.25+0.15+0.15=0.75$$

$$H_{T3}=\ p_4(1,3)+\ p_5(3,1)+\ p_7(2,3)+\ p_8(3,2)\ +\ p_9(3,3)$$
$$H_{T3}=0.06+0.06+0.15+0.15+0.09=0.51$$

And for the sample (1,1), the estimator

$$H_T(1,1)=\frac{30}{0.36}=83.33$$

for sample (1,2) the estimator

$$H_T(1,2)=\frac{30}{0.36}+\frac{90}{0.75}=203.33$$

and so on for samples(s); (2,1), (1,3), (3,1), (2,2), (2,3), (3,2) and (3,3) the estimators are 203.33, 191.18, 191.18, 120, 227.84, 227.84 and 107.84 respectively.

Now the mean $H_T=\ E(H_T)=\ \sum p(s)H_T(s)$

$$=\ (0.04)(83.33)+(0.1)(203.33)+(0.1)(203.33)+(0.06)(191.18)$$

$$+(0.06)(191.18) + (0.25)(120) + (0.15)(227.84) + (0.15)(227.84) + (0.09)(107.84)$$
$$= 175$$

and the variance

$$Var(H_T) = E[H_T - mean\ of\ H_T]^2$$

where $E(H_T) = \sum p(s)h_t(s)$, $h_t(s) = (H_T - mean\ of\ H_T)^2$

So,

$$Var(H_T) = (0.04)(83.33 - 175)^2 + (0.1)(203.33 - 175)^2 + (0.1)(203.33 - 175)^2$$
$$+ (0.06)(191.18 - 175)^2 + (0.06)(191.18 - 175)^2 + (0.25)(120 - 175)^2$$
$$+ (0.15)(227.84 - 175)^2 + (0.15)(227.84 - 175)^2 + (0.09)(107.84 - 175)^2$$
$$= 2527.88$$

Samples(s)	$p(s)$	x's	H_U	H_T
(1,1)	(0.2)(0.2) = 0.04	(30,30)	150	83.33
(1,2)	(0.2)(0.5) = 0.1	(30,90)	115	203.33
(2,1)	(0.5)(0.2) = 0.1	(90,30)	115	203.33
(1,3)	(0.2)(0.3) = 0.06	(30,55)	166.67	191.18
(3,1)	(0.3)(0.2) = 0.06	(55,30)	166.67	191.18
(2,2)	(0.5)(0.5) = 0.25	(90,90)	180	120
(2,3)	(0.5)(0.3) = 0.15	(90,55)	181.67	227.84
(3,2)	(0.3)(0.5) = 0.15	(55,90)	181.67	227.84
(3,3)	(0.3)(0.3) = 0.09	(55,55)	183.33	107.84
Mean			165	175
Variance			679.19	2527.88
			Hansen-Hurwitz	Horvitz-Thompson

Hansen-Hurwitz shows that if the data were collected from regions, groups or cluster of unequal sizes, it would be efficient to analyze the data using probability proportion to size

(compare to SRS). When the sample taken from the population is large the sampling with replacement is considered less accurate than sampling without replacement. But if the sample is small the probability of a particular unit to appear twice in the sample is also small and so both sampling with and without replacement is considered equivalent. The inclusion probability sampling improves the estimate by giving the larger unit higher chance to appear in the sample.

3. TEST OF INDEPENDENCE AND FITNESS

The complex surveys involve and contents various categories. Data is collected on each of those categories. The most important and concerning issues with different categorical or varieties of data is, if they are co-related or dependent. The test for dependence or independence checks if the given two categories or class of data changes with there is change in any one of them. Chi-square test of Independence is carried out when the sampling method is SRS, have two categorical variable and frequency count in the contingency table is greater than 5. If the sample has two variables, A and B, with levels a and b respectively, the general form of contingency table looks as below.

$Variable\ B$	$Variable\ A$				
	1	2	….	a	n_b
1	O_{11}	O_{12}	$O_{1.}$	O_{1a}	n_1
2	O_{21}	O_{22}	$O_{2.}$	O_{2a}	n_2
⋮	$O_{.1}$	$O_{.2}$	$O_{...}$	$O_{.a}$	n_3
b	O_{b1}	O_{b2}	$O_{b.}$	O_{ab}	n_b
n_a	n_1	n_2	$n_.$	n_a	n

Hypothesis can be defined as

H_0: Variable A and Variable B are independent

H_1: Variable A and Variable B are not independent

with degree of freedom

$$df = (a-1)(b-1) \quad \text{.....(3.1)}$$

Then the expected frequency:

$$E_{a,b} = (n_a * n_b)/n \text{.....(3.2)}$$

where n is total population of the sample.

Test statistics:

$$\chi^2 = \Sigma\left[\left(O_{a,b} - E_{a,b}\right)^2/E_{a,b}\right] \text{.....(3.4)}$$

If the value of test statistics, χ^2, is greater then the critical value, the null hypothesis H_0 is rejected.

Kish and Frankel (1974) had studied the effect of complex survey designs with stratification and clustering on statistics like subclass means, regression coefficients and correlation coefficients. They discovered the variance inflating effects of clustering was outweighed slightly by variance deflating effects of stratification. The ratio of the inflated variance of a statistics under a complex sampling scheme to the corresponding simple random sample variance is called a design effect (Kish 1965).

Hansen, Hurwitz and Madow (1953) formula for the design effect (deff) is given by,

$$deff = 1 + (s - 1)\rho \text{.....(3.5)}$$

where s is the average cluster size and ρ is the intra-cluster correlation coefficient (Rao, Thomas 1988). However this formula does not suffice when we move from inference on binary proportions to inference on multivariate data tables.

3.1. The K- Category Goodness-of-Fit Test

Let's consider one-way frequency table with the number of sample units in each of K categories of a discrete variable in domain U. Let n_U, represent the number of sample units in domain U, and let $n_1, n_2, \ldots\ldots, n_K$ represent the number of domain units in each of the K

categories. Then, $n_U = \Sigma n_k$. Let $p_1, p_2, \ldots\ldots, p_K$ be the proportion of units in each category in the finite population from which the sample is taken.

Simple hypothesis then can be tested for goodness-of-fit as;

$$H_0: p = p_k, \qquad k = 1,2, \ldots\ldots, K.$$
$$H_1 : p \neq p_k$$

If the data is treated as SRS the Pearson's statistics is given by,

$$X^2_{SRS} = \sum_{k=1}^{K} \frac{(n_k - n_U p_k)^2}{(n_U p_k)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1.1)$$

$$= n_U \sum_{k=1}^{K} \frac{(\hat{p}_k - p_k)^2}{p_k}$$

where $\hat{p}_k = n_k / n_U$ is the unweight proportion of sample units in category k of domain U. The standard test procedure rejects H_0 when,

$$X^2_{SRS} = \chi^2_{K-1}(\alpha) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1.2)$$

, the upper α percent point of χ^2_{K-1}. Unless the design is self-weighted, $\hat{p}_k$ above is not a consistent estimator of p_k. So χ^2_{SRS} will not be a useful measure of deviation of data from H_0 (Rao, Thomas 1988). Hence for general non-self-weighted designs, a weighted up form of Pearson's statistic is to be used, which is given by

$$X^2(G) = n_U \sum_{k=1}^{K} \frac{(\hat{p}_k - p_k)^2}{p_k} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1.3)$$

where $\hat{p}_k = \widehat{N}_k / \widehat{N}_U$ and $\hat{p}_k$ is a consistent estimator of p_k. Here, $\widehat{N}_U$ is the sum of sample weights of all units in domain U and $\widehat{N}_k$ is the sum of the weights of those units in U that are also in category k.

The likelihood ratio statistic G^2 is also used in categorical data analysis as an alternative to Pearson's χ^2 and is given by

$$G^2(G) = 2n_U \sum_{k=1}^{K} \hat{p}_k \ln(\hat{p}_k / p_k) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1.4)$$

$\chi^2(G)$ and $G^2(G)$ are both natural goodness-of-fit test statistics for complex design involving clustering. Their asymptotic distribution under H_0 will not be the familiar ${\chi^2}_{K-1}$, even for the self-weighting designs because of the non-zero intra-cluster correlation. Thus the rejection rule

$$X^2(G) = \chi^2_{K-1}(\alpha),$$

will not provide an α-level test. The asymptotic distribution of $\chi^2(G)$ and $G^2(G)$ under complex sample design have been obtained by Rao and Scott (1981) and provide the basis for their approach to the analysis of categorical data from complex surveys.

3.2. Alternatives to $\mathbf{X^2(G)}$ and $\mathbf{G^2(G)}$

Wald Tests: Let V represent the $(K-1)\times(K-1)$ covariance matrix of $\hat{p}$ and $\hat{V}$ be the estimate of V obtained from suitable method. Here $\hat{p} = (\hat{p}_1, \ldots\ldots, \hat{p}_{k-1})'$ represent the $(K-1)$ vector of estimated domain proportions with $\hat{p}_K = 1 - (\hat{p}_1, \ldots\ldots, \hat{p}_{k-1})$.

The Wald statistics is,

$$X^2_W(G) = (\hat{p} - p_0)' \hat{V}^{-1} (\hat{p} - p_0) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1.5)$$

where p_0 represent the corresponding $(K-1)$ vector of hypothesized proportions. This statistics is distributed asymptotically as χ^2_{K-1}under the null hypothesis and therefore provides an asymptotically exact α-level test when referred to $\chi^2_{K-1}(\alpha)$.

For the goodness of fit case, Thomas and Rao (1987) showed that the Wald test provides poor control of type I error unless the degree of freedom f for estimating $\hat{V}$ are much greater than the degree of freedom for the hypothesis. (Thomas and Rao 1988)

The F test,

$$F_W(G) = \frac{(f-K+2)}{f(K-1)} X^2_W(G) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1.6)$$

provides improved type I error control for the goodness-of-fit case.

Fay's jackknifed chi-square tests: This test provides a viable analysis strategy whenever suitable PSU (primary sample unit) or replicate data are available. (Thomas and Rao 1988)

One downside of Wald and Fay test procedure is that they require detailed survey information to estimate the covariance matrix V.

4. LOG- LINEAR MODEL

There are several statistical models that can be represented as generalized linear models. And among all those generalized linear models one of the cases is the log-linear model. Log-linear model analysis is the extended version of Pearson's chi-square test, which allows us to compare between more than two variables. Log-linear analysis does not differentiate between dependent and independent variables. It is the extended way of analyzing categorical data in 2×2 contingency table by taking the natural log of the cell-frequencies in the table. The following material in this section is adopted from *Discrete Multivariate Analysis: Theory and Practice*; Structural Models for Counted Data (Chapter 2) Bishop, Fienberg, and Holland; MIT; 1975.

4.1. Two-way 2×2 Table

Let's consider two way table with variables A and B each having two different (1 and 2) measurement and respective probabilities as follows;

		B		
		1	2	Total
A	1	p_{11}	p_{12}	p_{1+}
	2	p_{21}	p_{22}	p_{2+}
	Total	p_{+1}	p_{+2}	1

The log-linear model is written as ;

$$\log p_{ab} = u + u_{1(a)} + u_{2(b)} + u_{12(ab)} \quad \text{.....(4.1.1)}$$

$a = 1,2 \; and \; b = 1,2$ with constraints

$$\sum_a u_{1(a)} = \sum_b u_{2(b)} = \sum_a u_{12(ab)} = \sum_b u_{12(ab)} = 0 \text{.....(4.1.2)}$$

where if $l_{ab} = \log p_{ab}$ then the grand mean u is written as

$$u = \frac{l_{++}}{4} = \sum_{a,b} \frac{l_{ab}}{4} \text{.....(4.1.3)}$$

and the main effects $u_{1(a)}$, $u_{2(b)}$ and the interaction effect $u_{12(ab)}$ is written as

$$u_{1(a)} = \frac{l_{a+}}{2} - \frac{l_{++}}{4} \text{.....(4.1.4)}$$

$$u_{2(b)} = \frac{l_{+b}}{2} - \frac{l_{++}}{4} \text{.....(4.1.5)}$$

$$u_{12(ab)} = l_{ab} - \frac{l_{a+}}{2} - \frac{l_{+b}}{2} + \frac{l_{++}}{4} \text{.....(4.1.6)}$$

If we consider expected counts instead of probabilities the table becomes as follows;

		B		
		1	2	Total
A	1	n_{11}	n_{12}	N_1
	2	n_{21}	n_{22}	N_2
	Total	n_{+1}	n_{+2}	N

where $N = \sum_{a,b} n_{ab}$ and $n_{ab} = Np_{ab}$ and hence

$$\log n_{ab} = \log N + \log p_{ab} \text{.....(4.1.7)}$$

$$= u' + (u_{1(a)} + u_{2(b)} + u_{12(ab)})$$

where $u' = u + \log N$

If we define $l_{ab} = \log n_{ab}$ then the grand mean, main effects $u_{1(a)}$, $u_{2(b)}$ and interaction effect $u_{12(ab)}$ are given by the relations as described above.

4.2. Two-way $A \times B$ Table

Suppose we have a single sample of size N with A rows for variable A and B columns for variable B. Then for $\sum_{a,b} n_{ab} = N$ and $l_{ab} = \log n_{ab}$, $a = 1,2, \ldots, A$, $b = 1,2, \ldots, B$ the log-linear model is the same as

$$l_{ab} = u + u_{1(a)} + u_{2(b)} + u_{12(ab)} \text{.....(4.2.1)}$$

with constraints

$$\sum_a u_{1(a)} = \sum_b u_{2(b)} = \sum_a u_{12(ab)} = \sum_b u_{12(ab)} = 0 \text{.....(4.2.2)}$$

And we define grand mean,

$$u = \frac{l_{++}}{AB} \text{.....(4.2.3)}$$

main effect of variable A and B and interaction effect respectively,

$$u_{1(a)} = \frac{l_{a+}}{B} - \frac{l_{++}}{AB} \text{.....(4.2.4)}$$

$$u_{2(b)} = \frac{l_{+b}}{B} - \frac{l_{++}}{AB} \text{.....(4.2.5)}$$

$$u_{12(ab)} = l_{ab} - \frac{l_{a+}}{A} - \frac{l_{+b}}{B} + \frac{l_{++}}{AB} \ldots\ldots\ldots\ldots\ldots\ldots (4.2.6)$$

4.3. The $2 \times 2 \times 2$ table

$Variable\ C_1$

$Variable\ B$

$Variable\ A$		1	2
	1	n_{111}	n_{121}
	2	n_{211}	n_{221}

$Variable\ C_2$

$Variable\ B$

$Variable\ A$		1	2
	1	n_{112}	n_{122}
	2	n_{212}	n_{222}

We can describe 2×2 array by two separate log-linear model each for two different c's,

$$l_{abc} = v^{(c)} + v_{1(a)}^{(c)} + v_{2(b)}^{(c)} + v_{12(ab)}^{(c)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.3.1) \qquad a, b, c = 1,2$$

with $\sum_a v_{1(a)}^{(c)} = \sum_b v_{2(b)}^{(c)} = \sum_a v_{12(ab)}^{(c)} = \sum_b v_{12(ab)}^{(c)} = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.3.2)$

and grand mean, $u = \frac{1}{C}\sum_{c=1}^{C} v^{(c)}$, main effect and interaction effects,

$$u_{1(a)} = \frac{1}{C}\sum_{c=1}^{C} v_{1(a)}^{(c)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.3.3)$$

$$u_{2(b)} = \frac{1}{C}\sum_{c=1}^{C} v_{2(b)}^{(c)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.3.4)$$

$$u_{12(ab)} = \frac{1}{C}\sum_{c=1}^{C} v_{12(ab)}^{(c)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.3.5)$$

The deviations from these mean depend on the third variable and are defined as follows (Bishop et all 1975);

main effect of variable C,

$$u_{3(c)} = \frac{1}{C}\sum_{c=1}^{C} v^{(c)} - u \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.3.6)$$

interactions with variable C,

$$u_{13(ac)} = \frac{1}{C}\sum_{c=1}^{C} v_{1(a)}^{(c)} - u_{1(a)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.3.7)$$

$$u_{23(bc)} = \frac{1}{C}\sum_{c=1}^{C} v_{2(b)}^{(c)} - u_{2(b)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.3.8)$$

and three factor interaction effect,

$$u_{123(abc)} = \frac{1}{C}\sum_{c=1}^{C} v_{12(ab)}^{(c)} - u_{12(ab)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.3.9)$$

The one in all linear model for 2×2×2 table is ,

$$l_{abc} = u + u_{1(a)} + u_{2(b)} + u_{3(c)} + u_{13(ac)} + u_{12(ab)} + u_{23(bc)} + u_{123(abc)} \ldots\ldots\ldots (4.3.10)$$

with constraints

$$\sum_a u_{123(abc)} = \sum_b u_{123(abc)} = \sum_c u_{123(abc)} = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.3.11)$$

4.4. Models for four or more dimensions

Above we saw how a model for 3 dimensions, each with two categories, was written in terms of $v^{(c)}$ for each two dimension table defined by C categories of the third variable. And again the u-terms for the third variable were calculated from the averages across the two-dimension table. The deviations for third variables were then derived from these averages. For four dimension model with say P, Q, R and S categories, the model can be written in terms of $w^{(s)}$ for each three dimensions table with categories p, q and r and defined by S categories of the fourth variable. The u-terms for the third variable can be calculated from the averages across the

three-dimension table and the deviations for the fourth variable from these averages. For even higher dimension, Let's say h-dimension model, we go about the similar way.

4.5. The General Log-linear model for multi-way table

Let's consider a multi-way table with cross-classification of variables n_U survey samples in some domain U. Then the general log-linear model is of the form,

$$\log(p) = u(\theta)1 + X\theta \text{.....(4.5.1)}$$

where $\log(p)$ denotes a $S\ X\ 1$ vector of log probabilities with elements $\log(p_s)$, $s = 1,2,\dots,S$; 1 denotes a $S\ X\ 1$ vector of ones; X denotes a $S\ X\ r$ matrix of full column rank ($r \leq S-1$), with $X'1 = 0$ and θ is a $r\ X\ 1$ vector of parameters. $U(\theta)$ is a normalizing constant which ensures that $\sum p_k = 1$ and can be written as

$$U(\theta) = ln\left\{\frac{1}{1'\exp(X\theta)}\right\} \text{.....(4.5.2)}$$

Here $\exp(X\theta)$ is a $S\times 1$ vector having $\exp(x_k'\theta)$ as elements and x_k' is the kth row of design matrix X.

The following Pearson statistics is defined as a test for the specific design matrix X with rank $r < S$ and in terms of pseudo maximum likelihood estimators (MLE's), $\hat{p}$ and weighted-up proportion, $\widehat{\boldsymbol{P}}$,

$$X^2(L) = n_U \sum_{k=1}^{K} \left(\widehat{\boldsymbol{P}}_k - \hat{p}_k\right)^2 \Big/ \hat{p}_k \text{.....(4.5.3)}$$

And likelihood ratio statistics G^2 can be defined as,

$$G^2(L) = 2n_U \sum_{k=1}^{K} \hat{p}_k\, ln\left(\widehat{\boldsymbol{P}}_k \Big/ \hat{p}_k\right) \text{.....(4.5.4)}$$

In above equation, the notation S is the number of total cells in lexicographic order.

If we consider population samples A and the categories B then the log-likelihood ratio statistics is given by,

$$I(A,B) = \sum_{a=1}^{A}\sum_{b=1}^{B} x_{ab} lnx_{ab} - \sum_{a=1}^{A} x_{ab} lnx_{ab} - \sum_{b=1}^{B} x_{ab} lnx_{ab} + NlogN \quad \text{.....(4.5.5)}$$

Where x is the frequencies of the cell which can also be expressed as the proportions or probabilities and N is the observation total.[Kullback, Saloman; *Information Theory and Statistics*, New York; McGraw-Hill, 1950]

In the case of log-likelihood ratio statistics if the test holds one conclusion for A sample population over B categories then it holds the same test conclusion for any subset of population and categories. This is called a coherent property in the Simultaneous Test Procedure (STP) introduced by K.R. Gabriel. [Gabriel, K.R.; Simultaneous Test Procedure – Some Theory of Multiple Comparison; The Annals of Mathematical Statistics, Vol 40, No 1(Feb., 1969), pp 224-250]

The condition for the Simultaneous Test Procedure is

$$2I(A,B) > \epsilon \quad \text{.....(4.5.6)}$$

where ϵ is the upper α percentage point of the Chi-square distribution with $(A-1)(B-1)$ degrees of freedom.[Gabriel, K.R.; Simultaneous Test Procedure for Multiple Comparison on Categorical Data]

The test statistic $G^2(L)$ and $2I(A,B)$ is analogous and can be used for simultaneous test procedure.

If we consider $T, U\ and\ V$ as lexicographic numbers representing number of cells in the data such that $V < U < T$ then,

$$G^2(L_T) \geq G^2(L_U) \geq G^2(L_V) \quad \text{.....(4.5.7)}$$

And if we have $t = 1,2, \ldots, T$ then,

$$G^2(L) \geq \sum_{t=1}^{T} G^2(L_t) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.5.8)$$

where $t \leq T$.

$G^2(L_T)\ or\ G^2(L)$ is likelihood test statistics that checks the probability of a type I error at the upper α percentage point. Since $G^2(L_U),\ G^2(L_V)\ or\ \sum_{t=1}^{T} G^2(L_t)$ does not occur without $G^2(L_T)\ or\ G^2(L)$ occurring, following the coherent property of simultaneous test procedure, $G^2(L_U),\ G^2(L_V)\ or\ \sum_{t=1}^{T} G^2(L_t)$ test does not increase the probability of type I error beyond α.

5. THE RAO-SCOTT APPROACH

Rao-Scott first-order correction: A close first-order corrected test can be achieved by referring $X^2(G)/\delta$, the corrected Pearson statistics, to $\chi^2_{K-1}(\alpha)$ since the expected value of the asymptotic distribution of $X^2(G)$ is $\sum_{k=1}^{K-1}\delta$, so that $X^2(G)/\delta$, where

$$\delta = \frac{\sum_{k=1}^{K-1}\delta_k}{(K-1)} \text{.....(5.1)}$$

, has the same expected value as χ^2_{K-1}.

$$\widehat{\delta} = \frac{n_U}{K-1}\sum_{k=1}^{K}\frac{\hat{v}_{kk}}{p_k}$$

$$\widehat{\delta} = \frac{1}{K-1}\sum_{k=1}^{K}\frac{\hat{p}_k}{p_k}(1-\hat{p}_k)\hat{d}_k \text{.....(5.2)}$$

where $\hat{v}_{kk}$ represents the kth diagonal element of estimated covariance matrix $\hat{V}$ and

$$\hat{d}_k = \frac{\hat{v}_{kk}}{n_U^{-1}\hat{p}_k(1-\hat{p}_k)} \text{.....(5.3)}$$

is the estimated cell deff for the kth cell. First order corrected statistics do not involve n_U but depends on $\hat{p}_k$ (weighted estimates) and their estimated variance $\hat{v}_{kk}$,

$$X_c^2(G) = (K-1)\left(\sum_{k=1}^{K} \hat{v}_{kk}/p_k\right)^{-1}\left(\sum_{k=1}^{K} (\hat{p}_k - p_k)^2/p_k\right) \ldots\ldots\ldots\ldots\ldots (5.4)$$

Second-order correction: When $\hat{V}$, the full estimated covariance matrix, is known then an improved approximation of the asymptotic distribution of X^2 can be achieved using Satterthwaite approximation. The second-order correction can be written as,

$$X_S^2(G) = X^2(G)/\hat{\delta}_{.}(1 + \hat{a}^2)$$

$$X_S^2(G) = X_c^2(G)/(1 + \hat{a}^2) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.5)$$

under the null hypothesis as X_v^2, a chi-square random variable on $v = (K-1)/(1+\hat{a}^2)$ degrees of freedom with a corresponding form $G_S^2(G)$ for the likelihood ratio statistic. Here '$\hat{a}$', the coefficient of variation of the estimated δ_i's is given by

$$\hat{a}^2 = \frac{\sum_{k=1}^{K-1} \hat{\delta}_k^2}{[(K-1)\hat{\delta}_{.}^2] - 1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.6)$$

where $\hat{\delta}_k^2$ is the estimator of δ_k. (Thomas and Rao 1988)

Considering level of the second-order Satterthwaite approximation is approximately equal to the nominal α level, then actual level of the X^2 and X_c^2 tests can be approximated by

$$P\left(X^2 \geq \chi^2_{(K-1),\alpha}\right) = P\left(\chi_v^2 \geq \chi^2_{(K-1),\alpha}/[\hat{\delta}_{.}(1+\hat{a}^2)]\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.7)$$

and

$$P\left(X_c^2 \geq \chi^2_{(K-1),\alpha}\right) = P\left(\chi_v^2 \geq \chi^2_{(K-1),\alpha}/(1+\hat{a}^2)\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.8)$$

Residual Analysis

Standardized residuals can be defined as

$$\hat{e}_k = \frac{\hat{p}_k - p_k}{\hat{v}_{kk}^{1/2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.9)$$

$$= \frac{e_k}{\hat{d}_k^{1/2}} \text{ , } \quad k = 1,2, \ldots, K$$

here $\hat{e}_k$ are standardized residuals under SRS assumption

$$\hat{e}_k = \frac{\hat{p}_k - p_k}{\sqrt{\hat{p}_k(1-\hat{p}_k)/n_U}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.10)$$

Example on Goodness-of-fit (Thomas and Rao 1998)

Goodness of fit test also called Pearson's Chi-square goodness of fit test of a statistical data or model tells how well it fits the set of observation or observed data. Results of goodness of fit generally points out the differences in between the observed values and the expected values for the given model. The following example from the paper by Rao and Thomas (1988) is CCSS dataset containing variable, employment class, categorized by managements control degree and autonomy. Proportion estimated for employment class, used from the Canadian, U.S. and Swedish surveys, are given in Black and Myles (1986, Table 1B) for the domain of salaried employees. A test of homogeneity of proportions can be carried out between the Canadian and U.S. population to check if they are the same. In this example U.S. estimates were treated as fixed since the required design information on the U.S. survey was not available. Here the table below gives the estimated proportions ($\hat{p}_k$), Hypothesized proportions ($\hat{p}_k$), cell deffs ($\hat{d}_k$), and standardized residuals ($\hat{e}_k$) for the five categories of the employment class variable.

n_U = 1463	Estimated $\hat{p}_k$'s (Canada)	Hypothesized p_k's (U.S.)	Cell Deffs ($\hat{d}_k$'s)	Standardized Residuals ($\hat{e}_k$'s)
Decision-making managers	0.141	0.148	1.10	-0.72
Advisor-managers	0.034	0.052	1.96	-2.80
Supervisor	0.118	0.149	1.31	-3.2
Semi-autonomous workers	0.191	0.110	1.44	6.52
Workers	0.516	0.541	1.87	-1.35

Table 1: Estimated and Hypothesized Proportions, Deffs and Residuals for Categories of Employment Class

We have equations,

$$X^2(G) = n_U \sum_{k=1}^{K} (\hat{p}_k - p_k)^2 / p_k \quad \text{.....(5.11)}$$

And

$$\hat{\delta} = \frac{n_U}{K-1} \sum_{k=1}^{K} \frac{\widehat{v_{kk}}}{p_k} = \frac{1}{K-1} \sum_{k=1}^{K} \frac{\widehat{p_k}}{p_k} (1 - \widehat{p_k}) \widehat{d_k} \quad \text{.....(5.12)}$$

Using above equations $X^2(G)$ and $\hat{\delta}$ are computed to be 106.8 and 1.48 respectively with first order correction statistics, $X_c^2(G) = 72.3$. We can observe here that the value of $X^2(G)$ has reduced greatly by the first order correction. But referring to $\chi_4^2(0.05) = 9.45$ the goodness of fit test still provides strong evidence against null hypothesis (H_0) . Analyzing the table above one can tell the proportion of Decision-making managers and Workers in Canada and U.S. are almost the same and there are differences in the proportions of Semi-autonomous workers; higher in Canada and lower in U.S. Here the residual value of 1.35 shows that the difference in the proportions of workers between two countries is not significant.

The Balanced Repeated Replication (BRR) estimate of the variance of $\hat{p}$ is given by,

$$var(\hat{p}) = \frac{1}{K} \sum_{k=1}^{K} (\widehat{p_k} - \hat{p})^2 \quad \text{.....(5.13)}$$

Using the above BRR method of variance estimation we compute $\hat{V}$ to be,

$$\hat{V} = 10^{-5}\begin{pmatrix} 9.063 & 0.557 & -0.375 & -1.23 & -8.015 \\ 0.557 & 4.328 & 1.622 & 0.114 & -6.622 \\ -0.375 & 1.622 & 9.288 & -3.739 & -6.796 \\ -1.23 & 0.114 & -3.739 & 15.163 & -10.308 \\ -8.015 & -6.622 & -6.796 & 10.308 & 31.741 \end{pmatrix}$$

The second order Rao-Scott correction can be implemented by,

$$X_S^2(G) = \frac{X^2(G)}{\hat{\delta}(1+\hat{a}^2)} = \frac{X_c^2(G)}{(1+\hat{a}^2)} \quad \text{.....(5.14)}$$

The coefficient of variation $\hat{a}$ of the estimated δ_i's is given by,

$$\hat{a}^2 = \sum_{k=1}^{K} \frac{\hat{\delta}_k^2}{[(K-1)\hat{\delta}^2]-1} \quad \text{.....(5.15)}$$

Where $\hat{\delta}_k$ is the estimate of δ_k. Alternately $\hat{a}^2$ can be calculated without getting the individual eigenvalues, since $\sum \hat{\delta}_k^2$ can be expressed in closed form as below,

$$\sum_{k=1}^{K-1} \hat{\delta}_k^2 = n_U^2 \sum_{i=1}^{K} \sum_{j=1}^{K} \hat{v}_{ij}^2 / p_i p_j \quad \text{.....(5.16)}$$

Then we get $\hat{a}^2 = 0.182$ from above. Thus, $X_S^2(G) = 63.02$ on $\nu = 3.38$ degree of freedom ($\nu = (K-1)/(1+\hat{a}^2)$). From standard chi-square table, we find that $\chi_{3.38}^2(0.05) = 8.45$, which shows that the second order corrected test is highly significant. Assuming that the second order corrected test achieves a level equal to the normal level of 5%, the appropriate levels of the uncorrected X^2 and the first-order correction X_c^2 were obtained as 18% and 6% respectively from the following equations,

$$P\left(X^2 \geq \chi_{(K-1),\alpha}^2\right) = P\left(\chi_\nu^2 \geq \frac{\chi_{(K-1),\alpha}^2}{[\hat{\delta}(1+\hat{a}^2)]}\right) \quad \text{.....(5.17)}$$

And

$$P\left(X_C^2 \geq \chi_{(K-1),\alpha}^2\right) = P\left(\chi_\nu^2 \geq \frac{\chi_{(K-1),\alpha}^2}{(1+\hat{a}^2)}\right) \quad \text{.....(5.18)}$$

The correction is evidently valuable.

5.1. Test of homogeneity of Proportions

It is the comparison of proportions of categorical variable in independent samples. In complex survey the test of homogeneity can be carried out between two regional data within a single big survey or between two data of two big surveys. Here we consider a test of homogeneity of proportions of K-category variable across R regions (Rao and Thomas, 1988). The null hypothesis is given as

$$H_0 : p_1 = p_2 = \cdots = p_R \quad \text{.....(5.1.1)}$$

where $p_r, r = 1,2,\ldots,R$, denotes the $(K-1)$ vector of proportions in r region, with elements $p_{rk}, k = 1,2,\ldots,K-1$. A natural Pearson statistic for testing H_0 is given by

$$X^2(H) = n_U \sum_{r=1}^{R} \alpha_r \sum_{k=1}^{K} (\hat{p}_{rk} - \hat{p}_{+k})^2 / \hat{p}_{+k} \quad \text{.....(5.1.2)}$$

where $\alpha_r = \frac{n_U(r)}{n_U}, n_U(r)$ is the number of domain units in region r ,

n_U is the total domain size given by $\sum_{r=1}^{R} n_U(r)$,

$\hat{p}_{rk}$ is the estimate of p_{rk} based on weighted-up counts and

$\hat{p}_{+k} = \sum_{r=1}^{R} \alpha_k \hat{p}_{rk}$.

If $n_U(r)$ is unknown it is replaced by an estimate of its expected value given by $n(r)\widehat{N}_U(r)/\widehat{N}(r)$, where $\widehat{N}_U(r)$ and $\widehat{N}(r)$ are the sums of sample weights in region r of all units and of those units falling in domain U, respectively (Rao and Thomas, 1988).

Scott and Rao (1981) showed that the asymptotic distribution of $X^2(H)$, under the null hypothesis is of the form $\delta_1 W_1 + \delta_2 W_2 + \cdots + \delta_{(R-1)(K-1)} W_{(R-1)(K-1)}$, where W's are independent chi-square random variables, each on one degree of freedom. For the case of two regions, the δ's will be estimated by $\hat{\delta}_k, k = 1,2,\ldots,(K-1)$, the eigenvalues of an estimated generalized deff matrix given by

$$\hat{E}_H = n_U\left(\alpha_1(1-\alpha_1)\hat{P}_+^{-1}\right)\left(\hat{V}_1 + \hat{V}_2\right) \quad \text{.....(5.1.3)}$$

where $\hat{P}_+ = diag(\hat{p}_+) - \hat{p}_+\hat{p}_+'$ and $\hat{V}_1, \hat{V}_2$ are the estimated covariance matrices of $\hat{p}_1$ and $\hat{p}_2$ respectively.

For the general case $\hat{E}_H$ is given by

$$\hat{E}_H = n_U\left(A \otimes \hat{P}_+^{-1}\right) \hat{\Delta}_H \quad \text{....(5.1.4)}$$

where $A = diag(\alpha) - \alpha\alpha'$, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{R-1})'$,

and

$$\hat{\Delta}_H = \oplus_{r=1}^{R-1}\hat{V}_r + \hat{V}_R \otimes J_{R-1} \text{.....(5.1.5)}$$

Here $\hat{V}_r, r = 1,2, \ldots R$ are estimated covariance matrices for each independent region and J_{R-1} is an $(R-1)\times(R-1)$ matrix of ones (Rao and Thomas, 1988).

We can obtain first-order Rao-Scott corrections to $X^2(H)$. As a general case for R regions, the eigenvalues $\hat{\delta}$ is given by

$$\hat{\delta} = \frac{1}{(K-1)(R-1)} \sum_{r=1}^{R}(1-\alpha_r) \sum_{k=1}^{K} \frac{\hat{p}_{rk}}{\hat{p}_{+k}}(1-\hat{p}_{rk})\, \hat{d}_{r(k)} \text{ .....(5.1.6)}$$

where $\hat{d}_{r(k)} = n_U \hat{v}_{r(kk)} / \hat{p}_{rk}(1-\hat{p}_{rk})$ are the cell deffs for the kth region,

$\hat{v}_{r(kk)}$ being the kth diagonal element of $\hat{V}_r$.

The first-order corrected test is then given by

$$X_c^2(H) = \frac{X^2(H)}{\hat{\delta}}\ to\ \chi^2_{(R-1)(K-1)}(\alpha) \text{.....(5.1.7)}$$

The second-order corrected test is obtained by

$$X_S^2(H) = \frac{X_c^2(H)}{(1+\hat{a}^2)} \text{ .....(5.1.8)}$$

where

$$\hat{a}^2 = \sum_{k=1}^{K} \frac{\hat{\delta}_k^2}{\left[(K-1)\hat{\delta}^2\right]-1} \text{.....(5.1.9)}$$

with $\hat{\delta}$ given by above relation and $\sum \hat{\delta}_k^2$ replaced by $tr\hat{E}_H^2$ (sum of the diagonal elements of $\hat{E}_H^2$ and $(K-1)$ replaced by $(R-1)(K-1)$. The standardized residuals with an approximate standard normal distribution under H_0 are given by

$$\hat{e}_{rk} = \frac{(\hat{p}_{rk} - \hat{p}_{+k})}{\sqrt{var(\hat{p}_{rk} - \hat{p}_{+k})}} \text{ .....(5.1.10)}$$

where $r = 1,2,\ldots,R$ and $k = 1,2,\ldots,K$ and

$$var(\hat{p}_{rk} - \hat{p}_{+k}) =$$

$$\frac{1}{n_U^2}\left(\hat{p}_{+k}(1-\hat{p}_{+k})\right)\times\left\{\frac{n_U(n_U-2n_U(r))}{n_U(r)}\hat{d}_{r(k)} + \sum_{k=1}^{R} n_U(l)\hat{d}_{l(k)}\right\}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5.1.11)$$

Example: Test of homogeneity (two regions, i.e. R=2) (Rao and Thomas, 1988)

Below is the table with estimated proportions ($\hat{p}_r$), the cell deffs($\hat{d}_{(r)}$) and the standardized residuals($\hat{e}_r$) for both Eastern and Western provinces. Eastern Canada consists of survey regions Atlantic provinces, Quebec and Ontario and Western Canada consists of Prairie provinces, Alberta and British Columbia. Since design strata in the CCSS are subsets of the above six regions and since post stratification is carried out separately within these six regions the East and West constitutes two independent samples. Separate BRR variance estimations were carried out for East and West using seventeen design strata in each.

Table 2: Est Estimated Proportions, Deffs, and Residuals for Employment Class within Region

$n_U(1) = 1054$ $n_U(2) = 409$	Proportions $\hat{p}_r$, r=1,2		Regional Deffs $\hat{d}_{(r)}$,r=1,2		Std. Residuals $\hat{e}_r$, r=1,2	
	East	West	East	West	East	West
Decision-making managers	0.134	0.16	0.84	1.77	-1.01	1.01
Advisor-managers	0.033	0.035	2.37	0.91	-0.21	0.21
Supervisor	0.111	0.138	1.11	1.66	-1.18	1.18
Semi-autonomous workers	0.194	0.18	1.17	2.26	0.46	-0.46
Workers	0.528	0.487	1.38	3.34	0.83	-0.83

From the above table $X^2(H)$ and $\hat{\delta}$ were calculated as 4.47 and 1.76 from (5.1.2) and (5.1.6) respectively with first-order corrected statistics $X_c^2(H) = 2.54$ from (5.1.7). Comparing with $\chi_4^2(0.05) = 9.45$, $X^2(H) = 4.47$ and $X_c^2(H) = 2.54$ are not significant. Thus the hypothesis that there is no difference in the distribution

of employment class between Eastern and Western Canada cannot be rejected. Square of coefficient of variation, $\hat{a}^2$, of the eigenvalues of the deff matrix (5.1.3) was calculated to be 0.267 leading to the second-order corrected statistics $X_S^2(H)$ = 2.00. Referring to $\chi^2_{3.16}(0.05)$ = 8.08 (here degree of freedom $\nu = (K-1)/(1+\hat{a}^2)$= 3.16) $X_S^2(H)$ is not significant.

From (7) and (8) with $(K-1)$ replaced by $(R-1)(K-1)$ the actual level of $X^2(H)$ and $X_c^2(H)$, corresponding to a nominal 5% test, are approximately 25.6% and 6.5% respectively, which tells corrected tests are required for these data.

There are various ways of testing the homogeneity of the data set collected in surveys. Among many, Simultaneous Test Procedure (STP) is one that checks the homogeneity of larger data set and many smaller subsets contained in it. If A is the categories over which the survey is done and B is the regions over which its carried out then the log-likelihood ratio statistics $I(A,B)$ with degree of freedom; $df = (A-1)(B-1)$ and α as the error rate should satisfy the following for it to be considered as heterogeneous. (Gabriel, K.R; 1969)

$$2I(A,B) > \chi^2_{\alpha,df} \text{.....(5.1.12)}$$

Here

$$I(A,B) = I(A,B) = \sum_{a=1}^{A}\sum_{b=1}^{B} x_{ab} lnx_{ab} - \sum_{a=1}^{A} x_{ab} lnx_{ab} - \sum_{b=1}^{B} x_{ab} lnx_{ab} + NlogN \text{.....(5.1.13)}$$

(Kullback, Saloman; 1950).

Not only a big set of data, but several disjoint subset of data from a big set can be checked for homogeneity with STP. Let's consider k disjoint set combinations of original A categories and B regions data. Let the sub sets be $(A_1, B_1), (A_2, B_2), \dots, (A_k, B_k)$ of bigger set of data (A, B). Then if,

$$2\sum_{s=1}^{k} I(A_s, B_s) > \chi^2_{\alpha,df} \text{.....(5.1.14)}$$

where $\chi^2_{\alpha,df}$ is the chi-square statistics at type I error rate α with $df = (A-1)(B-1)$ degree of freedom, then at least one of those subset is heterogeneous. This way of checking tells, if the whole set of selected population is homogeneous or not.

5.2. Test of Independence

Let's consider domain U with n_U sample units with two-way cross-classified variables 1 and 2 with categories A and B respectively. Say, $p_{ab} = \frac{N_U(ab)}{N_U}$ be the abth cell proportions in domain U estimated by $\hat{p}_{ab} = \frac{\hat{N}_U(ab)}{\hat{N}_U}$, with domain total based on weighted-up counts. Let $\hat{p}_{a+} = \sum_{b=1}^{B} \hat{p}_{ab}\,, a = 1,2,\dots,A$ and $\hat{p}_{+b} = \sum_{a=1}^{A} \hat{p}_{ab}\,, b = 1,2,\dots,B$ be estimated row and column marginal proportions. The hypothesis of independence is as,

$$H_0: p_{ab} = p_{a+}p_{+b} \quad \text{.....(5.2.1)}$$

$a = 1,\dots,A \; and \; b = 1,\dots,B$

Considering estimated proportions $\hat{p}_{a+}$ and $\hat{p}_{+b}$, the Pearson X^2 statistic for testing H_0 is given by,

$$X^2(I) = n_U \sum_{a=1}^{A}\sum_{b=1}^{B} \frac{(\hat{p}_{ab}-\hat{p}_{a+}\hat{p}_{+b})^2}{\hat{p}_{a+}\hat{p}_{+b}} \quad \text{.....(5.2.2)}$$

Rao-Scott corrections to $X^2(I)$.

As earlier the first order correction to $X^2(I)$ can be obtained by dividing $X^2(I)$ by $\hat{\delta}$, where for two-way table

$$\hat{\delta} = \frac{1}{(A-1)(B-1)} \sum_{a=1}^{A}\sum_{b=1}^{B} \frac{\hat{p}_{ab}(1-\hat{p}_{ab})}{\hat{p}_{a+}\hat{p}_{+b}} \hat{d}_{ab} - \sum_{a=1}^{A}(1-\hat{p}_{a+})\hat{d}_{A(a)} - \sum_{b=1}^{B}(1-\hat{p}_{+b})\hat{d}_{B(b)} \quad \text{.....(5.2.3)}$$

and

$$\hat{d}_{ab} = estvar\,(\hat{p}_{ab}) / \left(n_U^{-1}\hat{p}_{ab}(1-\hat{p}_{ab})\right) \quad \text{.....(5.2.4)}$$

is the (i,j)th cell deff and,

$$\hat{d}_{A(a)} = estvar\,(\hat{p}_{a+})/\left(n_U^{-1}\hat{p}_{a+}(1-\hat{p}_{a+})\right)\text{.....(5.2.5)}$$

$$\hat{d}_{B(b)} = estvar\,(\hat{p}_{+b})/\left(n_U^{-1}\hat{p}_{+b}(1-\hat{p}_{+b})\right)\text{.....(5.2.6)}$$

are the deffs of the ath row and bth column margin, respectively.

So the first order correction is,

$$X_c^2(I) = \frac{X^2(I)}{\hat{\delta}}\text{.....(5.2.7)}$$

to $\chi^2_{(A-1)(B-1)}(\alpha)$.

Alternatively, $X^2(I)$ can also be expressed in terms of a vector $\hat{s}$ of residuals of length $(A-1)(B-1)$ as,

$$\hat{s}_{ab} = \hat{p}_{ab} - \hat{p}_{a+}\hat{p}_{+b}\text{.....(5.2.8)}$$

And

$$\hat{s} = \left(\hat{s}_{11}, \hat{s}_{12}, \dots, \hat{s}_{1,(B-1)}, \hat{s}_{21}, \dots, \hat{s}_{(A-1),1}, \hat{s}_{22}, \hat{s}_{23}, \dots, \hat{s}_{(A-1)(B-1)}\right)'$$

$\hat{P}_{A+} = \left(\hat{p}_{1+}, \dots, \hat{p}_{(A-1)+}\right)'$, vector of marginal proportion of length $(A-1)$

$\hat{P}_{+B} = \left(\hat{p}_{+1}, \dots, \hat{p}_{+(B-1)}\right)'$, vector of marginal proportion of length $(B-1)$

Then,

$$\hat{P}_A = diag\left(\hat{P}_{A+}\right) - \hat{p}_{A+}\hat{p}'_{A+}\text{.....(5.2.9)}$$

$$\hat{P}_B = diag\left(\hat{P}_{+B}\right) - \hat{p}_{+B}\hat{p}'_{+B}\text{.....(5.2.10)}$$

Now,

$$X^2(I) = n_U\hat{s}'\left(\hat{P}_A^{-1}\otimes\hat{P}_B^{-1}\right)\hat{s}\text{.....(5.2.11)}$$

The Pearson statistics $X^2(I)$ will have asymptotic distribution, as shown by Rao and Scott (1979,1981), as a weighted sums of

$$\delta_1 W_1 + \delta_1 W_2 + \cdots + \delta_{(A-1)(B-1)} W_{(A-1)(B-1)}$$

And δ's are estimated by the eigenvalues of the estimated design effects matrix $\widehat{D}_A = \left(\hat{P}_A^{-1} \otimes \hat{P}_B^{-1}\right)\hat{V}_s$ with $\hat{V}_s$ as the estimated covariance matrix of $\hat{s}$.

The second order correction is given by,

$$X_S^2(I) = \frac{X_c^2(I)}{(1+\hat{a})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.2.12)$$

Where

$$\hat{a}^2 = \sum_{q=1}^{(A-1)(B-1)} \frac{\hat{\delta}_q^2}{\left[(A-1)(B-1)\hat{\delta}^2\right]-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.2.13)$$

And

$$\sum_{q=1}^{(A-1)(B-1)} \hat{\delta}_q^2 = \sum_{a,a'}^{A-1} \sum_{b,b'}^{B-1} \frac{\hat{v}_{s,ab,a'b'}^2}{(\hat{p}_{a+}\hat{p}_{+b})(\hat{p}_{a'+}\hat{p}_{+b'})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.2.14)$$

Here $\hat{v}_{s,ab,a'b'}^2$ is the element of $\hat{V}_h$ corresponding to the covariance of $\hat{p}_{ab}$ and $\hat{p}_{a'b'}$. And $X_S^2(I)$ refers to $\chi_v^2(\alpha)$ with $v = (A-1)(B-1) \Big/ (1+\hat{a}^2)$

Example: Test of Independence (Rao,J.N.K. and D.R. Thomas;1988)

The following is the table with estimated cell proportions, $\hat{p}_{ab}$'s and corresponding design effects (deffs), $\hat{d}_{ab}$'s.

	Estimated Proportion ($\hat{p}_{ab}$'s)		Deffs of $\hat{p}_{ab}$'s ($\hat{d}_{ab}$'s)	
	Males	Females	Males	Females
Decision-making managers	0.103	0.038	1.20	1.31
Advisor-managers	0.018	0.016	0.74	1.95
Supervisors	0.075	0.043	1.81	0.92
Semi-autonomous workers	0.105	0.085	0.71	1.85
Workers	0.239	0.278	1.42	1.15
$n_U = 1463$				

Estimated Proportions of Deffs for Employment Class by Sex

Estimated marginal proportions $\hat{p}_{+b}$ and corresponding $\hat{d}_{B(b)}$ for the sex variable are (0.54, 0.46) and (1.29,1.29) respectively for males and females. The marginal deffs for employment class from goodness-of-fit example above was 1.53 and here it is 1.29, a smaller value. The unadjusted Pearson statistic, $X^2(I)$= 54.8, significant at 5% level (referring to $\chi_4^2(0.05) = 9.49$). From above table the average of eigenvalues of generalized deff matrix $\widehat{D}_A$ was calculated as $\hat{\delta} = 1.11$ leading to the corrected statistics $X_c^2(I) = 50.0$ which is significant with different smaller value then $X^2(I)$.

The value of second-order adjusted test was calculated as $X_S^2(I) = 38.4$ with $v = 3.07$ degrees of freedom, significant at 5% level.

6. EXAMPLES

Since the whole content of the paper is based on the attempt to understand the complexity of complex survey analysis, the data I have used here as for the examples is also from the complex survey carried out on the population throughout Canada (Statistics Canada(2004) Canadian Community Health Survey: Mental Health and Well Being; Catalogue no. 82-617-XIE). Although this survey involves population from ten provinces in five regions from the whole country, the actual number involved does not reflect the total number of present day population of Canada. For that reason the survey has weighted scale within it to represent the larger population. The number of units or cases in the survey is 22,191 and the weighted population is 15,840,147. The survey is conducted over various categories or variables including sex, age, employment, parental status, education, income etc. For my purpose of analysis the example I used here has categories of different medication taken by units across regions or provinces.

6.1. Goodness-of-Fit Test

First and fore most we start with the classical goodness-of-fit test with our sets of survey categories. Here I have selected six different categories of medication taken (A) across five regions(B) of Canada. The survey had responses like yes, no, not sure, null etc. for each categories but I am using only the 'yes' responses as frequencies in each categories across different regions. The size of the sample population (n) in this example is 6,254. The first line in each row for the six different categories in the following Table 1 is the actual or observed responses $(O_{a,b})$ and the second line are the expected values $(E_{a,b})$.

Table 1

Categories(A) ↓ Provinces(B) →	Maritimes	Quebec	Ontario	Prairies	BC	Total
Took medication to help sleep - 12 mo	427	324	900	488	325	2464
	455.84	359.71	862.83	471.60	314.01	
Took medication/reduce anxiety - 12 mo	315	252	469	244	148	1428
	264.18	208.47	500.05	273.32	181.98	
Took mood stabilizers - 12 mo	41	58	117	70	58	344
	63.64	50.22	120.46	65.84	43.84	
Took anti-depressants - 12 mo	354	256	629	346	234	1819
	336.52	265.55	636.97	348.15	231.81	
Took med. treat psychotic behav. - 12mo	15	18	46	28	23	130
	24.05	18.98	45.52	24.88	16.57	
Took stimulants - 12 mo	5	5	29	21	9	69
	12.77	10.07	24.16	13.21	8.79	
Total	1157	913	2190	1197	797	6254

Following are mathematical relations used in this example to acquire the desired values and hence the result.

Hypothesis to be tested;

H_0: Variable A and Variable B are independent

H_1: Variable A and Variable B are not independent

Expected frequencies $E_{a,b} = (n_a * n_b)/n$……(6.1.1)

where n_a is the total observed frequency for category a, n_b is the total observed frequency for region b, n is the total sample population i.e. $n = 6{,}254$ in this example and $a = 1,2, \dots 6; b = 1,2, \dots 5$.

Degree of freedom$(df) = (a-1)(b-1)$

$= 20$

Test statistics, $\chi^2 = \Sigma\left[\left(O_{a,b} - E_{a,b}\right)^2 / E_{a,b}\right]$……..(6.1.2)

$$\chi^2 = 72.9924567$$

From Chi-Square table the critical value $\chi^2_{0.05,20} = 31.41$, which is less than 72.99 and hence the null hypothesis; variable A and variable B are independent, is rejected. That is to say

the test shows some dependencies between the regions and the responses collected within those regions.

6.2. Calculation of Design Effect

As we all know the collection of data in surveys involves different approaches of getting responses from each unit or case. Those approaches and ways of doing things have certain effect on the type of data collected which then affect the outcome of the survey analysis. Rao and Scott considered those affects called design effects in analyzing the complex survey data. Here is an attempt to show what the design effect are in mathematical calculation and how they are achieved in a simple way.

Here again in Table 2, I have used the same data as in Table 1. The first row in each category is the actual or observed response($O_{a,b}$) for that category in different regions. The second row for each category is the proportion($p_{a,b}$) corresponding to the observed response ($O_{a,b}$) in each region and the total observed response(n_a) in that category.

$$i.e.\, p_{a,b} = \frac{o_{a,b}}{n_a} \text{.....(6.2.1)}$$

Here, $p_{1,1} = \frac{427}{2464} = 0.101497504$

$p_{1,2} = \frac{324}{2464} = 0.09878049$

$p_{1,3} = \frac{900}{2464} = 0.1159495$

$p_{1,4} = \frac{488}{2464} = 0.1078691$ and so on.

Table 2

Categories with their units/cases count along the regions with their respective proportions ***($p_{a,b}$).***

Categories(A) ↓ Regions(B) →	**Maritimes**	**Quebec**	**Ontario**	**Prairies**	**BC**	**Total**
Took medication to help sleep - 12 mo	427	324	900	488	325	2464
Proportion($p_{a,b}$)	0.101497504	0.098780488	0.115949498	0.107869142	0.134854772	
$Sqrs = (p_{a,b} - p_{a\mu})^2$	9.17485E-05	0.000151181	2.37505E-05	1.02843E-05	0.000565428	
Took medication/reduce anxiety - 12 mo	315	252	469	244	148	1428
Proportion($p_{a,b}$)	0.074875208	0.076829268	0.060422572	0.053934571	0.061410788	
$Sqrs = (p_{a,b} - p_{a\mu})^2$	0.000110283	0.000155143	1.56108E-05	0.000108974	8.77837E-06	
Took mood stabilizers - 12 mo	41	58	117	70	58	344
Proportion($p_{a,b}$)	0.009745662	0.017682927	0.015073435	0.015473033	0.02406639	
$Sqrs = (p_{a,b} - p_{a\mu})^2$	3.31973E-05	4.73305E-06	1.883E-07	1.17908E-09	7.32568E-05	
Took anti-depressants - 12 mo	354	256	629	346	234	1819
Proportion($p_{a,b}$)	0.084145472	0.07804878	0.081035816	0.07648099	0.097095436	
$Sqrs = (p_{a,b} - p_{a\mu})^2$	4.60421E-06	1.561E-05	9.2913E-07	3.04565E-05	0.00022788	
Took med. treat psychotic behav. - 12m	15	18	46	28	23	130
Proportion($p_{a,b}$)	0.003565486	0.005487805	0.005926308	0.006189213	0.009543568	
$Sqrs = (p_{a,b} - p_{a\mu})^2$	5.26637E-06	1.38785E-07	4.35127E-09	1.08155E-07	1.35661E-05	
Took stimulants - 12 mo	5	5	29	21	9	69
Proportion($p_{a,b}$)	0.001188495	0.00152439	0.00373615	0.00464191	0.00373444	
$Sqrs = (p_{a,b} - p_{a\mu})^2$	3.69406E-06	2.51571E-06	3.91451E-07	2.34525E-06	3.89313E-07	
Total	4207	3280	7762	4524	2410	22183

And the third row for each category is the squares,

$$Sqrs = (p_{a,b} - p_{a\mu})^2 \text{.....(6.2.2)}$$

where, $p_{a\mu} = \frac{n_a}{n}$

So, $p_{1\mu} = 0.111076$, $p_{2\mu} = 0.0643736$ and so on.

Then the sum of square $\sum(p_{a,b} - p_{a\mu})^2$ is the sum for all the squares in each category along all regions.

For first category

$$\sum(p_{a,b} - p_{a\mu})^2 = 0.00084239$$

For the second category

$$\sum(p_{a,b} - p_{a\mu})^2 = 0.00039879$$

and so on.

Table 3

Categories with their respective proportion mean($p_{a\mu}$), sum of squares $\sum(p_{a,b} - p_{a\mu})^2$, SRS Variance Cluster Variance and Design Effect (DEFF).

	$p_{a\mu}$	$\sum(p_{a,b} - p_{a\mu})^2$	Variance(SRS)	Cluster Variance	DEFF
Took medication to help sleep - 12 mo	0.111076049	0.000842392	4.45127E-06	5.26495E-05	11.82795852
Took medication/reduce anxiety - 12 mo	0.064373619	0.000398789	2.71525E-06	2.49243E-05	9.179393061
Took mood stabilizers - 12 mo	0.015507371	0.000111377	6.88256E-07	6.96104E-06	10.11402772
Took anti-depressants - 12 mo	0.08199973	0.00027948	3.39355E-06	1.74675E-05	5.147264454
Took med. treat psychotic behav. - 12m	0.005860344	1.90838E-05	2.62645E-07	1.19274E-06	4.54124872
Took stimulants - 12 mo	0.00311049	9.33579E-06	1.3979E-07	5.83487E-07	4.174032354

$$Variance(SRS) = \frac{p_{a\mu}(1-p_{a\mu})}{(n_a-1)} \text{ .....(6.2.3)}$$

and,

$$Cluster\ Variance = \frac{\sum(p_{a,b}-p_{a\mu})^2}{(B-1)} \text{ .....(6.2.4)}$$

where B is the total number of regions. i.e. 5 in this case. And finally

$$the\ Design\ Effect(DEFF) = \frac{Cluster\ Variance}{Variance(SRS)} \text{ .....(6.2.5)}$$

Here in this example we get DEFF values 11.82795852, 9.179393061, 10.11402772, 5.147264454, 4.54124872 and 4.174032354 for categories one to six respectively. These values of design effect are later used for Rao-Scott first order correction.

6.3. Rao-Scott First order correction

Now since we have gone through the process of obtaining design effects, it is now possible to calculate the first-order correction for Rao-Scott approach. Here again in Table 4 below we have the same $K = 1,2,\ldots,6$ categories. But since we are using one way frequency, the region in consideration is only one. For our purpose I am using the proportional weight($\hat{p}$) of Maritimes province.

Here $n_U = 1157$ is the total sample unit in Maritimes province with 'Yes' response. And $n_1 = 427,\ n_2 = 315,\ n_3 = 41,\ n_4 = 354,\ n_5 = 15,$ and $n_6 = 5$ is the number of total cases in each of the K categories such that $n_U = \sum_{k=1}^{K} n_k,\ k = 1,2,\ldots,6$. The proportion $\hat{p}_k,\ k = 1,2,\ldots,6$ for each category is calculated as $\hat{p}_k = \frac{n_k}{n_U}$. The estimated proportion p_k is obtained from the data considering the total cases for each categories over all regions and the total survey 'Yes' response for all six categories. Design effect for each category due to clustering is given under $\widehat{\boldsymbol{d}}_{\boldsymbol{k}}$ in the table below. These $\widehat{\boldsymbol{d}}_{\boldsymbol{k}}$ values are carried from the previous example above.

Table 4
Population proportion ($\hat{\boldsymbol{p}}_{\boldsymbol{k}}$), EstimatedProportion($\boldsymbol{p}_{\boldsymbol{k}}$) and respective design effect($\widehat{\boldsymbol{d}}_{\boldsymbol{k}}$) for each categories in Maritime region.

n_U=1157	$\hat{\boldsymbol{p}}_{\boldsymbol{k}}$	$\boldsymbol{p}_{\boldsymbol{k}}$	$\widehat{\boldsymbol{d}}_{\boldsymbol{k}}$
Took medication to help sleep - 12 mo	0.369057908	0.393987848	11.82795852
Took medication/reduce anxiety - 12 mo	0.272255834	0.228333866	9.179393061
Took mood stabilizers - 12 mo	0.035436474	0.055004797	10.11402772
Took anti-depressants - 12 mo	0.305963699	0.290853854	5.147264454
Took med. treat psychotic behav. - 12mo	0.012964564	0.020786697	4.54124872
Took stimulants - 12 mo	0.004321521	0.011032939	4.174032354

Then the hypothesis to be tested for goodness-of-fit is as;

$$H_0: p = p_k , \qquad k = 1,2, \dots ,6.$$

$$H_1 : p \neq p_k$$

If the data is treated as SRS the Pearson's statistics is given by,

$$X^2 = \sum_{k=1}^{K} \frac{(n_k - n_U p_k)^2}{(n_U p_k)} \qquad \text{(6.3.1)}$$

$$X^2 = n_U \sum_{k=1}^{K} \frac{(\hat{p}_k - p_k)^2}{p_k}$$

$$= 28.69227856$$

The critical value $\chi^2_{0.05,5} = 11.07$ is for degree of freedom $(K-1) = 5$ and confidence level 5%, which is less than 28.69 so we reject the null hypothesis.

Now with First-order correction (Rao-Scott,1979),

$$X_C^2 = \frac{X^2}{\hat{\delta}} \qquad \text{(6.3.2)}$$

where,

$$\hat{\delta} = \frac{1}{K-1} \sum_{k=1}^{K} \frac{\hat{p}_k}{p_k} (1 - \hat{p}_k) \hat{d}_k \qquad \text{(6.3.3)}$$

$$\hat{\delta} = 5.88$$

Hence,

$$X_C^2 = \frac{X^2}{\hat{\delta}} = \frac{28.69}{5.88}$$

$$X_C^2 = 4.879$$

The first-order correction gives much smaller statistic than the original Pearson's statistics. The corrected one is less than $\chi^2_{0.05,5} = 11.07$ and so we can't reject the null hypothesis. This implies that the corrected statistics tell the conclusion of the analysis has to be changed. The previous non-corrected statistics gave a conclusion that the proportion of data collected in Maritimes region are not same as the average proportions of data collected through-out other regions. But the corrected statistics tell that after taking design effect into consideration the proportion of data collected in Maritimes region are as same proportion to the average proportions of data collected through-out other regions.

6.4. Simultaneous Test Procedure

Simultaneous Test procedure (STP) test the statistics if A set of population sample over B categories or combination statistics of smaller set of population from bigger set over smaller set of categories from bigger set is heterogeneous.

The table below has population samples from ten provinces is same six categories as above. For Simultaneous Test Procedure, we are going to take the log likelihood ratios of the sample data.

In general, if we consider the categories (A) and population samples (B) then the log likelihood ratio statistics is given by (Kullback, Solomon 1959),

$$I(A,B) = \sum_{a=1}^{A}\sum_{b=1}^{B} x_{ab} ln x_{ab} - \sum_{a=1}^{A} x_{ab} ln x_{ab} - \sum_{b=1}^{B} x_{ab} ln x_{ab} + NlogN \ldots\ldots\ldots\ldots (6.4.1)$$

where x_{ab} ,$A = 6$, $a = 1,2, \ldots 6$, $B = 10$, $b = 1,2, \ldots ,10$ is the frequency from corresponding cell and $N = 6254$ is the grand total.

Here population samples (B) are from the Canadian provinces Newfoundland and Labrador(NL), Prince Edward Island (PEI), Nova Scotia (NS), New Brunswick (NB), Quebec (QC), Ontario (ON), Manitoba (MB), Saskatchewan (SK), Alberta (AB) and British Columbia (BC)

Table 5

Frequency Table for Categories (A) along the provinces (B).

Categories(A) ↓ Provinces(B) →	NL	PEI	NS	NB	QC	ON	MB	SK	AB	BC	Total
Took medication to help sleep - 12 mo	72	50	177	128	324	900	124	108	256	325	2464
Took medication/reduce anxiety - 12 mo	67	30	146	72	252	469	75	56	113	148	1428
Took mood stabilizers - 12 mo	6	4	21	10	58	117	20	19	31	58	344
Took anti-depressants - 12 mo	67	35	178	74	256	629	101	76	169	234	1819
Took med. treat psychotic behav. - 12m	2	2	7	4	18	46	10	2	16	23	130
Took stimulants - 12 mo	1	1	1	2	5	29	3	6	12	9	69
Total	215	122	530	290	913	2190	333	267	597	797	6254

For simplicity and easy calculation the above equation for log likelihood ratio statistics is broken down into following equations,

$$L_A(B) = x_{Ab} lnx_{Ab} - \sum_{b=1}^{B=10} x_{Ab} lnx_{Ab} \text{.....(6.4.2)}$$

$$L_B(A) = x_{aB} lnx_{aB} - \sum_{a=1}^{A=6} x_{aB} lnx_{aB} \text{.....(6.4.3)}$$

$$L_b(A) = x_{Ab} lnx_{Ab} - \sum_{a=1}^{A=6} x_{ab} lnx_{ab} \text{.....(6.4.4)}$$

$$L_a(B) = x_{aB} lnx_{aB} - \sum_{b=1}^{B=10} x_{ab} lnx_{ab} \text{.....(6.4.5)}$$

And finally,

$$I(A,B) = L_A(B) - \sum_{a=1}^{A=6} L_a(B) \text{.....(6.4.6)}$$

or

$$I(A,B) = L_B(A) - \sum_{b=1}^{B=10} L_b(A) \text{.....(6.4.7)}$$

The following is the table with the calculated natural log for the values in Table 5 and the values for all above equations.

Natural Log table for Table 5

Provinces	NL	PEI	NS	NB	QC	ON	MB	SK	AB	BC	Total	$L_a(B)$
Took medication to help sleep - 12 mo	307.92	195.60	916.18	621.06	1872.96	6122.16	597.71	505.67	1419.57	1879.74	19242.71	4804.14
Took medication/reduce anxiety - 12 mo	281.71	102.04	727.61	307.92	1393.42	2884.63	323.81	225.42	534.19	739.59	10373.04	2852.70
Took mood stabilizers - 12 mo	10.75	5.55	63.93	23.03	235.51	557.17	59.91	55.94	106.45	235.51	2009.18	655.43
Took anti-depressants - 12 mo	281.71	124.44	922.36	318.50	1419.57	4053.36	466.13	329.14	866.95	1276.55	13653.49	3594.80
Took med. treat psychotic behav. - 12m	1.39	1.39	13.62	5.55	52.03	176.12	23.03	1.39	44.36	72.12	632.78	241.81
Took stimulants - 12 mo	0.00	0.00	0.00	1.39	8.05	97.65	3.30	10.75	29.82	19.78	292.15	121.43
Total	1154.69	586.09	3324.62	1644.27	6223.68	16844.73	1934.11	1491.80	3815.97	5324.64	54666.07	$L_A(B)$ 12321.47
$L_b(A)$	271.20	157.08	680.93	366.83	1242.16	2953.64	460.22	363.49	814.63	1101.37	$L_B(A)$ 8462.72	

From the table above we have the following values, $\mathrm{L_A(B)} = 12321.47$ and $\mathrm{L_B(A)} = 8462.72$

Then,

$$I(A,B) = L_A(B) - \sum_{a=1}^{A=6} L_a(B)$$

$$= 12321.47 - (4804.14 + 28.52.7 + 655.43 + 3594.8 + 241.81 + 121.43)$$

$$= 51.176$$

and,

$$2I_5(A,B) = 102.35$$

or

$$I(A,B) = L_B(A) - \sum_{b=1}^{B=10} L_b(A)$$

$$= 8462.72 - (271.2+157.08+680.93+366.83+1242.16+2953.64+$$

460.22+363.49+814.63+1101.37)

$$= 51.176$$

$$2I_5(A,B) = 102.35$$

Now the critical value at 5% point of chi-square distribution ($type\ I\ error\ rate\ at\ \alpha = 0.05$) with degree of freedom (df); $(A-1)(B-1) = (6-1)(10-1) = 45$, is , $\chi^2_{0.05,45} = 57.505$. Comparing $\chi^2_{0.05,45}$ to $2I(A,B)$ we can see the $2I(A,B)$ is larger, implying the total set of population over the ten provinces is heterogeneous on the surveyed six categories.

To check STP on smaller sets of population and categories I have calculated the corresponding statistics for each case below.

Table 6 has the data from three provinces; ON, AB and BC, deleted.

Log likelihood Statistics for Table 6 is;

$$2I_6(A,B) = 49.44$$

and

$$\chi^2_{0.05,30} = 43.773$$

And again $2I(A,B) > \chi^2_{0.05,30}$, and hence a heterogeneous set.

Table 6

Frequency Table for Categories (A) *along the provinces* (B).

Provinces	NL	PEI	NS	NB	QC	ON	MB	SK	AB	BC	Total
Took medication to help sleep - 12 mo	72	50	177	128	324		124	108			983
Took medication/reduce anxiety - 12 mo	67	30	146	72	252		75	56			698
Took mood stabilizers - 12 mo	6	4	21	10	58		20	19			138
Took anti-depressants - 12 mo	67	35	178	74	256		101	76			787
Took med. treat psychotic behav. - 12m	2	2	7	4	18		10	2			45
Took stimulants - 12 mo	1	1	1	2	5		3	6			19
Total	215	122	530	290	913		333	267			2670

Table 7 has the data from four provinces; NS, ON, AB and BC, deleted.

Log likelihood Statistics for Table 7 is,

$$2I_7(A,B) = 37.23$$

and

$$\chi^2_{0.05,25} = 37.652$$

Here we see $2I(A,B) < \chi^2_{0.05,30}$, and hence not a heterogeneous set. The population over the province NS made the above set of population in Table 6 a heterogeneous one.

Table 7
Frequency Table for Categories (A) *along the provinces* (B).

Provinces	NL	PEI	NS	NB	QC	ON	MB	SK	AB	BC	Total
Took medication to help sleep - 12 mo	72	50		128	324		124	108			806
Took medication/reduce anxiety - 12 mo	67	30		72	252		75	56			552
Took mood stabilizers - 12 mo	6	4		10	58		20	19			117
Took anti-depressants - 12 mo	67	35		74	256		101	76			609
Took med. treat psychotic behav. - 12m	2	2		4	18		10	2			38
Took stimulants - 12 mo	1	1		2	5		3	6			18
Total	215	122		290	913		333	267			2140

Table 8 has the data from four provinces; QC,ON, SK and BC, and two categories; 'Took mood stabilizers - 12 mo' and 'Took stimulants - 12 mo' deleted.
Log likelihood Statistics for Table 8 is,

$$2I_8(A,B) = 35.6$$

and

$$\chi^2_{0.05,15} = 24.996$$

And here again we see $2I(A,B) > \chi^2_{0.05,30}$, and hence a heterogeneous set. Data in Table 7 is not heterogeneous and the data in Table 8 is heterogeneous. Deleting the data from two categories made the set a heterogeneous.

Table 8

Frequency Table for Categories (A) *along the provinces* (B).

Provinces	NL	PEI	NS	NB	QC	ON	MB	SK	AB	BC	Total
Took medication to help sleep - 12 mo	72	50	177	128			124		256		807
Took medication/reduce anxiety - 12 mo	67	30	146	72			75		113		503
Took mood stabilizers - 12 mo											
Took anti-depressants - 12 mo	67	35	178	74			101		169		624
Took med. treat psychotic behav. - 12m	2	2	7	4			10		16		41
Took stimulants - 12 mo											
Total	208	117	508	278			310		554		1975

Table 9 has the data from one province ON and categories 'Took medication/reduce anxiety - 12 mo', 'Took anti-depressants - 12 mo' and 'Took stimulants - 12 mo', deleted.

Log likelihood Statistics for Table 9 is,

$$2I_9(A,B) = 23.4$$

and

$$\chi^2_{0.05,16} = 26.296$$

And here again we see $2I(A,B) < \chi^2_{0.05,30}$, and hence not a heterogeneous set.

Table 9

Frequency Table for Categories (A) *along the provinces* (B).

Provinces	NL	PEI	NS	NB	QC	ON	MB	SK	AB	BC	Total
Took medication to help sleep - 12 mo	72	50	177	128	324		124	108	256	325	1564
Took medication/reduce anxiety - 12 mo											
Took mood stabilizers - 12 mo	6	4	21	10	58		20	19	31	58	227
Took anti-depressants - 12 mo											
Took med. treat psychotic behav. - 12m	2	2	7	4	18		10	2	16	23	84
Took stimulants - 12 mo											
Total	80	56	205	142	400		154	129	303	406	1875

Table 10 has the data from provinces ON, MB, SK, AB and BC deleted.

Log likelihood Statistics for Table 10 is,

$$2I_{10}(A,B) = 27.37$$

and

$$\chi^2_{0.05,20} = 28.412$$

And here again we see $2I(A,B) < \chi^2_{0.05,30}$, and hence not a heterogeneous set.

Table 10

Frequency Table for Categories (A) along the provinces (B).

Provinces	NL	PEI	NS	NB	QC	ON	MB	SK	AB	BC	Total
Took medication to help sleep - 12 mo	72	50	177	128	324						751
Took medication/reduce anxiety - 12 mo	67	30	146	72	252						567
Took mood stabilizers - 12 mo	6	4	21	10	58						99
Took anti-depressants - 12 mo	67	35	178	74	256						610
Took med. treat psychotic behav. - 12m	2	2	7	4	18						33
Took stimulants - 12 mo	1	1	1	2	5						10
Total	215	122	530	290	913						2070

Table 11 has the data from provinces NL, PEI, NS, NB and QC deleted.

Log likelihood Statistics for Table 11 is,

$$2I_{11}(A,B) = 20.68$$

and

$$\chi^2_{0.05,20} = 28.412$$

And here again we see $2I(A,B) < \chi^2_{0.05,30}$, and hence not a heterogeneous set.

Table 11

Frequency Table for Categories (A) *along the provinces* (B).

Provinces	NL	PEI	NS	NB	QC	ON	MB	SK	AB	BC	Total
Took medication to help sleep - 12 mo						900	124	108	256	325	1713
Took medication/reduce anxiety - 12 mo						469	75	56	113	148	861
Took mood stabilizers - 12 mo						117	20	19	31	58	245
Took anti-depressants - 12 mo						629	101	76	169	234	1209
Took med. treat psychotic behav. - 12m						46	10	2	16	23	97
Took stimulants - 12 mo						29	3	6	12	9	59
Total						2190	333	267	597	797	4184

Here from above we see

$$2I_5(A,B) = 102.35 > \chi^2_{0.05,45}$$

but

$$2I_{10}(A,B) + 2I_{11}(A,B) = 27.37 + 20.68 = 48.05 < \chi^2_{0.05,45}$$

So, none of the disjoint sets of data in Table 10 and 11 are heterogeneous.

STP allows us to test the homogeneity or heterogeneity of various sub samples regions under a complex survey. It tells us if the sample we are testing is worth considering based on its homogeneity.

7. CONCLUSION

The techniques and procedures of analysis of surveys and samples have evolved greatly compared to what it was in the very early days. Nonetheless the concept leading to the present day understanding of surveys and its analytic conclusion has the seed of all those basic statistical concepts used in the very beginning. In this paper I started with the very concept of what are the sampling techniques with the examples to calculate statistical values from them. Then the simple and popular way of data analysis using goodness of fit test, homogeneity and independence was discussed. Those test showed the nature of data used for the analysis. The examples I have included in this paper at last on those test was to illustrate how those test were carried out. The concept of log-linear and model based on it was discussed to show the theory behind it. Many statistical software used in the field carry out the log-linear analysis in simple easy steps. But the user doesn't see the mathematical theory behind it. The discussion in this paper tries to show log-linear models are taken into account for two-way, three way and multi-way table arising from complex surveys.

Complex surveys have many levels of complexities. One is the way data is collected. Collection of data involves different techniques of sampling as discussed in chapter 2 and those give rise to the several errors eventually leading to a wrong conclusion of analysis of those data. To correct those errors Rao-Scott came up with some correction considering the design effect of the data collection. The examples with the theory on Rao-Scott correction have shown that the data do need some correction sometimes to have true analysis.

Then there are efficient ways of testing big complex data for their consistency using simultaneous test procedure (STP). The concept and examples in this paper has tested the efficiency of STP.

There are several concepts and ideas associated with the topics I have included in this paper. There are cases where Rao-Scott correction work, there are other things to consider while talking those concepts but it's the level of research that limit me from going too far and beyond the deep complexity. I have taken and meant the very simplest of the cases in every concepts and theory I have talked about in this paper. There are definitely more to be studied and further to be researched. It's a continuous and ongoing process of my research that will continue forward with refinements and improvements in my understanding in coming days.

REFERENCES

- Scheaffer, R. L., W. Mendenhall and R. Lyman Ott.1979. *Elementary Survey Sampling.* Cengage Learning, pp 59.
- Kish, L., and M. R. Frankel 1974. "Inference from Complex Samples"(with discussion), *Journal of the Royal Statistical Society*, ser. B, 36: 1-37.
- Kish, L. 1965. *Survey Sampling*. New York: Wiley.
- Hansen, M. H., W. H. Hurwitz and W. G. Madow (1953). *Sample Survey Methods and Theory*. Vol 2. New York: Wiley.
- Rao,J. N. K. and D. Roland Thomas; "The Analysis of Cross-Classified Categorical Data from Complex Sample Surveys."; *Sociological Methodology* Vol 18(1988), pp 213-269.
- Rao, J. N. K. and D. Roland Thomas. 1987. "Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics Under Cluster Sampling*." Journal of the American Statistical Association* 82:630-36.
- Bishop, Y. M. M., S. E. Fienberg and P. W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press.
- Kullback, Saloman; *Information Theory and Statistics*, New York; McGraw-Hill, 1950.
- Gabriel, K.R.; "Simultaneous Test Procedure – Some Theory of Multiple Comparison."; *The Annals of Mathematical Statistics*, Vol 40, No 1(Feb., 1969), pp 224-250.
- Gabriel, K.R.; "Simultaneous Test Procedure for Multiple Comparison on Categorical Data."; *Journal of the American Statistical Association*
  Vol. 61, No. 316 (Dec., 1966), pp. 1081-1096.
- Black, D., and J. Myles. 1986."Dependent Industrialization and the Canadian Class Structure: A Comparative Analysis of Canada, the United States, and Sweden." *Canadian Review of Sociology and Anthropology* 23:157-81.

- Rao, J. N. K., and A. J. Scott. 1979. “The Analysis of Categorical Data from Complex Sample Survey: Chi-Squared Test for Goodness of Fit and Independence in Two-Way Tables.” *Journal of the American Statistical Association* 76:221-30.
- Statistics Canada (2004) Canadian Community Health Survey: Mental Health and Well Being; Catalogue no. 82-617-XIE.